NAVAL RESEARCH LOGISTICS QUARTERLY

MARCH 1972
VOL. 19, NO. 1

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGIS

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this quarterly are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

THE RELIABILITY OF MULTICOMPONENT SYSTEMS SUBJECT TO CANNIBALIZATION

Richard M. Simon

National Institutes of Health
Bethesda, Maryland

ABSTRACT

A reliability model for multicomponent multistate systems is presented. This is a generalization of a model previously studied by Hirsch, Meisner, and Boll. In the earlier model, when a failure occurs for which no replacement spare is available, the locations using the same type of part as that having failed are "cannibalized" so as to allocate the shortages to locations where they are least detrimental to system performance. Here, we permit certain restrictions to be imposed upon the cannibalization procedure, and develop effective techniques for relating the probability laws governing the level of system performance to the system structure, cannibalization policy, kit of spare parts, and part reliabilities.

1. INTRODUCTION

Hirsch, Meisner, and Boll introduced in [3] a reliability model of multicomponent, multistate systems. Their model is novel in the sense that a component failure for which no spare part replacement is available is serviced by performing a cannibalization within the system. Thus, shortages are allocated to locations so as to minimize their degrading effect on system performance. The only restriction on the cannibalization process is that associated with each location of the system is a part type, and no part types are substitutes for each other. Thus, part interchanges can take place only among locations associated with the same part type. The authors of [3] show that under suitable conditions the probability laws governing the level of system performance can be efficiently calculated.

In the study [3], it was assumed that complete interchangeability is permitted among the locations associated with the same part type. In real systems, this assumption is not always met. In submarine systems, for example, there are many instances in which a part type has many different applications within the system, and the maintenance policy does not permit complete interchangeability among these applications [1]. For example, if two locations are in different equipments, each of which is difficult to access, then interchangeability between the locations may not be permitted. In some circumstances, different operating environments give rise to one-way restrictions; that is, it may be permitted to cannibalize a part from location λ to location λ', but not from λ' to λ. In this paper we shall permit such interchangeability restrictions, and show that under suitable conditions the probability laws governing the level of system performance can be estimated. The generalization presented here was motivated by participation in a study of the Polaris submarine system [1]. For complex kinds of interchangeability restrictions such as found in real systems, the interval estimates of reliability presented here are obtained in a more computationally efficient manner than can be obtained by computer simulation.

It is assumed that initially there are a finite number of spares of each part type, and that no more spares become available. After each failure for which a spare replacement is available, the failed part is discarded and the spare installed. The states of the locations determine the state of the system by

means of the system structure function. For an $M+1$ state system, the system state takes values on the set $0, 1, \ldots, M$. State M is interpreted as perfect system performance and state 0 represents complete system failure. If a part failure occurs for which no spare replacement is available, the system is cannibalized so as to maximize the system state subject to the constraints imposed by the interchangeability restrictions.

Our objective is to relate the probability laws governing the level of system performance to the system structure function, the cannibalization policy, the kit of spares available, and the component reliabilities. The following assumptions are made in this model:

1. Failures are detected instantaneously, and part replacements and interchanges are performed instantaneously.

2. Part lifetimes are independent random variables, and parts of the same type have identical lifetime distributions which are not affected by cannibalization.

Though the first assumption may seem unreasonable, it was motivated by the desire to study the "material readiness" of the system; that is, to focus attention on the relationship between the kit of spares available and system performance [1]. The second assumption is also an approximation to reality, but in most systems failure data are collected on a part-type basis and hence a more realistic assumption would not be of use. Both assumptions considerably simplify the mathematical theory. In this paper, we restrict our attention to "admissible" cannibalization policies [3]; that is, policies which cannibalize after each unspared failure to maximize the state of the system subject to the interchangeability restrictions. The notion of "admissible" cannibalization policy was introduced in [3], and in the model of [3] all admissible policies were equally good in a probabilistic sense. In the model presented here, if there are one-way cannibalization restrictions, then the admissible policies are not equally good. Admissible policies which interchange parts in such a way as to leave most flexibility after future failures are in a probabilistic sense better than those which do not. This topic is studied in [5], but in this paper the class of admissible policies will be treated as a unit.

The principal result of this paper is an efficient method of bounding the probability laws governing the level of system performance as a function of time. In order to obtain these bounds, certain separability conditions on the system structure function and the interchangeability restrictions must be assumed. These conditions are called the "minimum condition" and the "overlap condition". A system may be represented graphically as in Figure 1. The vertices represent locations, and the system state is the number of distinct arrows which can be reached from the left-most (root) vertex by passing through only vertices representing locations containing operable parts. It is shown in [4] that the minimum and overlap conditions are satisfied if the system is represented as in Figure 1 by a symmetric series-parallel graph in which interchangeability is permitted among all vertices on the same level (the same distance from the left-most vertex).

2. DEFINITIONS

We shall let n denote the number of locations in the system and B^n denote the set of binary n-tuples. The state of the locations at any time will be specified by a binary n-tuple v. $v_j = 1$ means that the jth location contains an operable part. The state of the system is specified by the states of the locations through the *system structure function* $\phi : B^n \to \{0, 1, \ldots, M\}$. The structure function is assumed to have the properties $v \geq v' \Rightarrow \phi(v) \geq \phi(v')$, $\phi(\underline{0}) = 0$, and $\phi(\underline{1}) = M$, where $\underline{0}$ and $\underline{1}$ denote vectors of all zeros and all ones, respectively.

The interchangeability restrictions are specified by a *restriction* mapping μ where for each location

λ, $\mu(\lambda)$ is the set of locations into which it is permitted to cannibalize from λ. We assume that all elements of $\mu(\lambda)$ are associated with the same part type, and that if $\lambda' \epsilon \mu(\lambda)$ and $\lambda'' \epsilon \mu(\lambda')$, then $\lambda'' \epsilon \mu(\lambda)$. As a convention it is assumed that $\lambda \epsilon \mu(\lambda)$ for any λ. It is not assumed that $\lambda' \epsilon \mu(\lambda)$ implies that $\lambda \epsilon \mu(\lambda')$; that is, one-way interchangeability restrictions are permitted. Two locations are said to belong to the same *communicating class* if parts can be exchanged between them in both directions. A communicating class, s, is said to be *closed* if and only if $\lambda \epsilon s$ and $\lambda' \epsilon \mu(\lambda)$ imply that $\lambda' \epsilon s$. Thus, a part can never be cannibalized from a location within a closed communicating class for installation outside of the class. A communicating class is said to be *isolated* if $\lambda \epsilon s$, and $\lambda \epsilon \mu(\lambda')$ imply $\lambda' \epsilon s$. A part can never be cannibalized from a location outside of an isolated communicating class for installation within the class. A communicating class may be both, either, or neither closed and/or/nor isolated. In the model of [3], the set of locations associated with the same part type form a single communicating class.

We shall let $e_k \epsilon B^n$ denote the unit vector with 1 in the kth position and 0 elsewhere. A *basic cannibalization* is a transformation $T: B^n \to B^n$ such that if $T(v) = v'$, then there exist indices i, j (not necessarily distinct) such that $\lambda_i \epsilon \mu(\lambda_j)$ and $v' - v = e_i - e_j$ (where λ_i and λ_j denote the ith and jth locations, respectively). Thus, a basic cannibalization represents the operation of taking a good part from one location and placing it in a location which does not contain a good part, such that the interchange is permitted by the restriction mapping. For $v \epsilon B^n$ let $[v]$ denote the set of $v' \epsilon B^n$ such that for some integer k, $v' = T_k(T_{k-1} \ldots (T(v)) \ldots)$ where $T_1, \ldots, T_k$ are basic cannibalizations. Thus $[v]$ denotes the set of v' that can be obtained from v by a sequence of basic cannibalizations. A *cannibalization* is a mapping $T: B^n \to B^n$ such that $T(v) \epsilon [v]$ for all $v \epsilon B^n$. It should be noticed that a cannibalization is a policy which specifies the interchanges to be performed given the states of the locations. This definition of cannibalization is a direct generalization of that given in [3].

A cannibalization, T, applied to a binary vector, v, effects a change in system state from $\phi(v)$ to $\phi T(v)$. The function ϕT describes the system after cannibalization and is called the *cannibalized structure function* [3]. A cannibalization, T, is called *admissible* [3] if $\phi T(v) = \max_{v' \epsilon [v]} \phi(v')$ for all $v \epsilon B^n$. In general, there may be many points in $[v]$ at which ϕ will take on the maximum value, and hence there are many admissible cannibalizations. It follows from the definition of admissible that for any two admissible cannibalizations T and T', $\phi T = \phi T'$.

For any subset s of locations, the mapping $\pi_s: B^n \to B^n$ is defined by

$$(\pi_s(v))_j = \begin{cases} v_j & \text{if } \lambda_j \epsilon s \\ 1 & \text{otherwise,} \end{cases}$$

for each location λ_j (where $(x)_j$ denotes the jth component of the vector x). Thus, π_s is a mapping which "turns on" the locations outside of s while leaving the locations of set s unchanged. We shall let Q_i denote the set of locations associated with part type i, and let N denote the number of part types. We shall use the notation $\pi_i = \pi_{Q_i}$.

These definitions can be illustrated with the aid of Figure 1. The vertices represent locations, and the number inside a vertex represents the part type associated with the location. The symbol above a vertex represents the location number of the vertex. The value of the structure function is given by the number of distinct arrows which can be reached in the graph from the leftmost vertex by passing through only locations containing operable parts. Thus, $\phi(v) = v_1 v_2 (v_4 v_8 + v_5 v_9 - v_4 v_8 v_5 v_9) v_{12}$

$(v_{14}+v_{15})+v_1v_3(v_6v_{10}+v_7v_{11}-v_6v_{10}v_7v_{11})v_{13}(v_{16}+v_{17})$. Suppose that the restriction mapping is given by; $\mu(\lambda_1)=\mu(\lambda_2)=\mu(\lambda_3)=\{\lambda_1, \lambda_2, \lambda_3\}$, $\mu(\lambda_4)=\mu(\lambda_5)=\mu(\lambda_6)=\mu(\lambda_7)=\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7\}$, $\mu(\lambda_{14})=\mu(\lambda_{15})=\mu(\lambda_{16})=\mu(\lambda_{17})=\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7, \lambda_{14}, \lambda_{15}, \lambda_{16}, \lambda_{17}\}$, $\mu(\lambda_8)=\mu(\lambda_9)=\mu(\lambda_{10})=\mu(\lambda_{11})=\{\lambda_8, \lambda_9, \lambda_{10}, \lambda_{11}\}$, $\mu(\lambda_{12})=\mu(\lambda_{13})=\{\lambda_{12}, \lambda_{13}\}$. Thus, the communicating classes are $S_1=\{\lambda_1, \lambda_2, \lambda_3\}$, $S_2=\{\lambda_4, \lambda_5, \lambda_6, \lambda_7\}$, $S_3=\{\lambda_{14}, \lambda_{15}, \lambda_{16}, \lambda_{17}\}$, $S_4=\{\lambda_8, \lambda_9, \lambda_{10}, \lambda_{11}\}$, and $S_5=\{\lambda_{12}, \lambda_{13}\}$. Class S_1 is closed, but not isolated; class S_3 is isolated, but not closed; class S_2 is neither closed nor isolated; and classes S_4 and S_5 are closed and isolated.

3. CHARACTERIZATION OF THE STRUCTURE FUNCTION

3.1 Basic Theory

A structure function ϕ is said to satisfy the *minimum condition* [3] if and only if $\phi=\min_{1\leq i\leq N}\phi\pi_i$. (This functional notation is to be interpreted as asserting that for all $v\epsilon B^n$, $\phi(v)=\min_{1\leq i\leq N}\phi\pi_i(v)$.) A cannibalized structure function ϕT is said to satisfy the minimum condition if and only if $\phi T=\min_{1\leq i\leq N}\phi T\pi_i$. Theorem 2 to follow shows that if ϕT satisfies the minimum condition, then for all $v\epsilon B^n$,

$$\phi T(v)=\min_{1\leq i\leq N}\phi\pi_i T(v). \tag{3.1}$$

Thus, the minimum condition can be interpreted as asserting that a single part type is responsible for the value assumed by the cannibalized structure function in the sense that the value remains unchanged if all inoperable parts of other types suddenly become operable. Consider for example the system represented graphically in Figure 2. Clearly, $\phi(v)=v_1v_2v_3+v_1v_4v_5$. Suppose that the restriction mapping is given by: $\mu(\lambda_1)=\mu(\lambda_2)=\mu(\lambda_4)=\{\lambda_1, \lambda_2, \lambda_4\}$, $\mu(\lambda_3)=\mu(\lambda_5)=\{\lambda_3, \lambda_5\}$. It is easily shown that for any admissible cannibalization T, the cannibalized structure function, ϕT satisfies the minimum condition.

Actually, the minimum condition for a cannibalized structure function is somewhat stronger than the assertion of expression (3.1). Given only the validity of expression (3.1) for a cannibalized

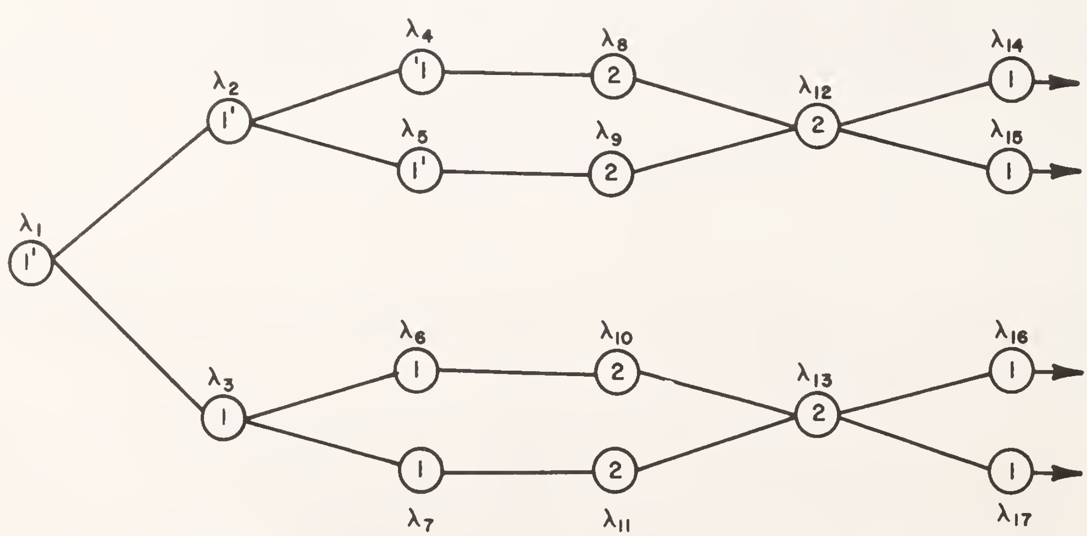


FIGURE 1.

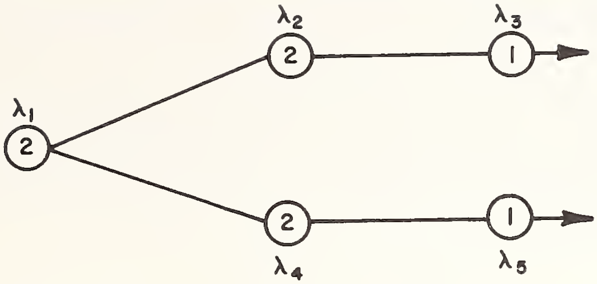


FIGURE 2.

structure function ϕT, it would in general be true that $\phi T(V) < \min_{1 \leq i \leq N} \phi T \pi_i(v)$. Consider, for example, the system represented in Figure 3. The structure function for this system is $\phi(v) = v_5 \cdot (2v_1 + 2v_2 \cdot v_4 + 2v_3 \cdot v_4 - 2v_2 \cdot v_3 \cdot v_4)$. Suppose that the restriction mapping is given by: $\mu(\lambda_1) = \mu(\lambda_2) = \{\lambda_1, \lambda_2\}$, $\mu(\lambda_3) = \{\lambda_3\}$, $\mu(\lambda_4) = \{\lambda_4\}$, $\mu(\lambda_5) = \{\lambda_5\}$. Consider the binary state $v = (0, 1, 0, 1, 1)$. Then $\phi T(v) = 2$ for any admissible cannibalization T, so we can define $T(v) = v$. It is thus seen that $\min_{1 \leq i \leq 3} \phi \pi_i T(v) = \phi \pi_1 T(v) = 2$, although $\min_{1 \leq i \leq 3} \phi T \pi_i(v) = 4$. The definition of T can be extended so that T is admissible,

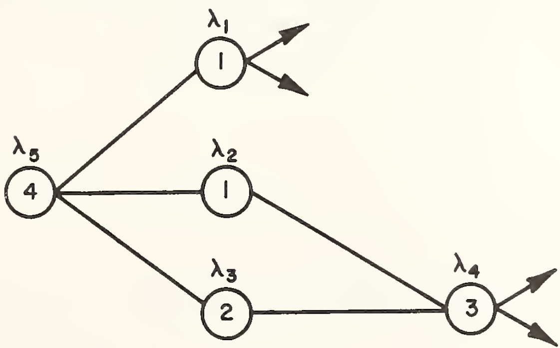


FIGURE 3.

and so the expression (3.1) holds for all binary states. As seen above, however, ϕT will not satisfy the minimum condition.

The following result is easily established.

THEOREM 1: If there exists an admissible cannibalization, T, such that ϕT satisfies the minimum condition, then $\phi T'$ satisfies the minimum condition for any admissible cannibalization T'.

The results presented in Theorem 2 were established in [3] with proofs which are valid for the more general restriction mappings considered here.

THEOREM 2: If ϕT satisfies the minimum condition, then for all $v \epsilon B^n$

(a) $$\min_{1 \leq i \leq N} \phi T \pi_i = \min_{1 \leq i \leq N} \phi \pi_i T, \qquad \text{and}$$

(b) $$\phi T(v) = \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' \mid \phi T \pi_i(v') \geq k\}}(v)$$

where $I_A(\ )$ denotes the indicator function for set A.

In [3], $w_i(v)$ was defined as the number of operable parts of type i when the state of the locations is v. For any given structure, the *associated unrestricted structure* is defined here as that system

having the same locations and structure function as the original system but complete interchangeability is permitted among the locations associated with the same part type. In the model of [3], a structure is its own associated unrestricted structure. If T is an admissible cannibalization on an unrestricted structure, then $\phi T\pi_i(v)$ depends only on $w_i(v)$. $n_i(k)$ was defined in [3] as the minimum value of $w_i(v)$ such that $\phi T\pi_i(v) \geqslant k$. It was shown in [3] that in an unrestricted structure if T is admissible and ϕT satisfies the minimum condition, then

$$\phi T = \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' \mid w_i(v') \geqslant n_i(k)\}}. \tag{3.2}$$

We shall use this result to obtain an upper bound on ϕT for structures which have interchangeability restrictions.

LEMMA 1: If T and T^* denote admissible cannibalizations on a system and its associated unrestricted structure respectively, then $\phi T \leqslant \phi T^*$.

PROOF: It is easily seen that T is a cannibalization on the associated unrestricted structure, hence the result follows from the admissibility of T^*.

The following result is derived directly from Lemma 1 and expression (3.2).

THEOREM 3: If T and T^* are admissible cannibalizations on a system and its associated unrestricted structure, and if ϕT^* satisfies the minimum condition, then

$$\phi T \leqslant \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' \mid w_i(v') \geqslant n_i(k)\}}$$

We shall now focus attention on obtaining a lower bound on ϕT. Let T be an admissible cannibalization. $\tilde{n}_i(k)$ is defined as the minimum l, such that for any $v\epsilon B^n$, if $w_i(v) \geqslant l$ then $\phi T\pi_i(v) \geqslant k$. It is important to recognize that in general $\phi T\pi_i(\quad)$ is not a function of $w_i(v)$ alone, but depends also upon the distribution among the communicating classes of the parts of type i.

The following example shows that even if $w_i(v) \geqslant \tilde{n}_i(k)$ for all part-types i, it is still possible that $\phi T(v) < k$ for T, an admissible cannibalization. Consider the system depicted in Figure 2. The structure function is $\phi(v) = v_1 \cdot v_2 \cdot v_3 + v_1 \cdot v_4 \cdot v_5$. Suppose that the restriction mapping is given by: $\mu(\lambda_1) = \{\lambda_1\}$, $\mu(\lambda_2) = \{\lambda_1, \lambda_2\}$, $\mu(\lambda_3) = \{\lambda_3\}$, $\mu(\lambda_4) = \{\lambda_1, \lambda_4\}$, $\mu(\lambda_5) = \{\lambda_5\}$. Thus $\tilde{n}_1(1) = 1$ and $\tilde{n}_2(1) = 2$. Consider $v = (1, 0, 1, 1, 0)$. For all admissible cannibalizations T, $T(v) = v$, and $\phi T(v) = 0$, although $w_1(v) = \tilde{n}_1(1)$, and $w_2(v) = \tilde{n}_2(1)$.

The following condition on the cannibalized structure function is in the same spirit as the minimum condition, and will lead to an easily evaluated lower bound on ϕT. Let T be an admissible cannibalization. For any $v\epsilon B^n$ and for $i = 1, 2, \ldots, N$ let $K_i(v) = \max\ \{k \mid w_i(v) \geqslant \tilde{n}_i(k)\}$. Then ϕT is said to satisfy the *overlap condition*, if and only if for all $v\epsilon B^n$ $\min_{1 \leqslant i \leqslant N} K_i(v) \leqslant \phi T(v)$. The overlap condition has the following interpretation. Consider any $v\epsilon B^n$ and suppose that we are told only how many parts of each type are operable. Then, in general, we cannot know with certainty the system state after cannibalization ($\phi t3v$)), because we do not know where the shortages are. $K_i(v)$ denotes the highest system state which we can be sure of for $T\pi_i(v)$ given only our information. The overlap condition asserts that we can be sure that $\phi T(v)$ is at least as great as the minimum of these highest assured values. The overlap condition is similar to the minimum condition, but less restrictive, as the following theorem demonstrates.

THEOREM 4: Let T be an admissible cannibalization. If ϕT satisfies the minimum condition, then ϕT satisfies the overlap condition.

PROOF: Consider any $v \epsilon B^n$. Since ϕT satisfies the minimum condition, $\phi T')v) = \min_{1 \leq i \leq N} \phi T \pi_i(v)$. By definition of $K_i(v)$, $\phi T \pi_i(v) \geq K_i(v)$. Hence $\phi T(v) \geq \min_{1 \leq i \leq N} K_i(v)$.

The following example indicates a simple case in which ϕT satisfies the overlap condition, but not the minimum condition. The structure function for Figure 4 is $\phi(v) = v_9(v_1 v_5 v_3 v_7 + v_2 v_6 v_4 v_8)$. Suppose that the restriction mapping is given by:

$$\mu(\lambda_9) = \{\lambda_9\},\ \mu(\lambda_1) = \{\lambda_1\},\ \mu(\lambda_2) = \{\lambda_2\},\ \mu(\lambda_3) = \mu(\lambda_4) = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\},\ \mu(\lambda_5) = \{\lambda_5\},\ \mu(\lambda_6) = \{\lambda_6\},\ \mu(\lambda_7) = \mu(\lambda_8) = \{\lambda_5, \lambda_6, \lambda_7, \lambda_8\}.$$

Consider $v = (1, 0, 0, 1, 0, 1, 0, 1, 1)$. For any admissible cannibalization T, $\phi T(v) = \phi T \pi_2(v) = \phi T \pi_3(v) = 1$, but $\phi T(v) = 0$. Hence ϕT does not satisfy the minimum condition. It can be shown, however, that ϕT satisfies the overlap condition. It can be seen that $n_1(2) = n_2(2) = 4$, $n_1(1) = 3$, $n_3(1) = n_3(2) = 1$. Thus for the specified v, $K_1(v) = K_2(v) = 0$.

THEOREM 5: Let T be an admissible cannibalization. If ϕT satisfies the overlap condition, then

$$\phi T(v) \geq \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | w_i(v') \geq \tilde{n}_i(k)\}}(v).$$

PROOF: Consider any $v \epsilon B^n$.

$$\phi T(v) = \sum_{k=1}^{M} I_{\{v' | w_i(v') \geq k\}}(v).$$

But since ϕT satisfies the overlap condition, $\phi T(v) \geq \min_{1 \leq i \leq N} K_i(v)$.

Hence

$$\{v' \mid \phi T(v') \geq k\} \supset \{v' \mid \min_{1 \leq i \leq N} K_i(v) \geq k\}.$$

Thus

$$\phi T(v) \geq \sum_{k=1}^{M} I_{\{v' | \min_{1 \leq i \leq N} k_i(v) \geq k\}}(v)$$

$$= \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | K_i(v') \geq k\}}(v).$$

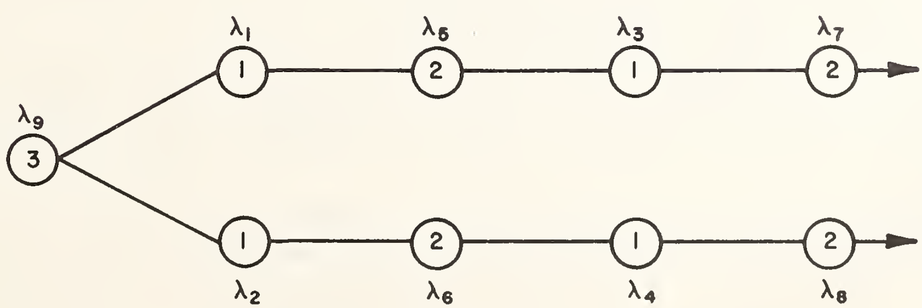


FIGURE 4.

But by the definitions of $\tilde{n}_i(k)$ and $K_i(v')$, $K_i(v') \geqslant k$ if and only if $w_i(v') \geqslant \tilde{n}_i(k)$. Hence

$$\phi T(v) \geqslant \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | w_i(v') \geqslant r_i(k)\}}(v).$$

The following result is a combination of Theorems 3 and 5.

THEOREM 6: Let T and T^* denote admissible cannibalizations on a system and its associated unrestricted structure respectively. If ϕT and ϕT^* satisfy the overlap condition and minimum condition respectively, then

$$\sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | w_i(v') \geqslant \tilde{n}_i(k)\}} \leqslant \phi T \leqslant \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | w_i(v') \geqslant n_i(k)\}},$$

Thus, for a certain class of structures, Theorem 6 provides an upper and lower bound on $\phi T(v)$ which depend only upon the number of operational parts of each type.

3.2 Calculation of $\tilde{n}_i(k)$

The evaluation of the constants $n_i(k)$ for use in expression (3.2) is fairly straightforward. The constants $\tilde{n}_i(k)$ are, however, of a more complex nature as they involve the cannibalization restrictions. In this section we wish to indicate that under certain conditions the $\tilde{n}_i(k)$ can be calculated in a simple manner.

Analogous to the definition of $w_i(v)$ provided in [3], for any set s of locations we define $w_s(v)$ as the number of operable locations in s when the states of the locations are specified by v. Thus, $w_i(v) = w_{Q_i}(v)$. For any given structure, the *associated closed isolated structure* is defined by making each communicating class both closed and isolated. If T' is an admissible cannibalization on the associated closed isolated structure, then for any communicating class s, $\phi T' \pi_s(v)$ depends only on $w_s(v)$. We can therefore define $n_s(k)$ as the minimum value of $w_s(v)$ such that $\phi T' \pi_s(v) \geqslant k$. Thus, $n_i(k) = n_{Q_i}(k)$. The symbol T' will be used consistently to denote any admissible cannibalization on the associated closed isolated structure.

A system is said to satisfy the *minimum condition within* Q_i if for all $v \epsilon B^n$, $\phi T' \pi_i(v) = \min \phi T' \pi_s(v)$, where the minimum is taken over the communicating classes of Q_i. It was shown in [5] that if a system satisfies the minimum condition within Q_i, then $\phi T \pi_i(v) = \min \phi \pi_s T \pi_i(v)$, where T is an admissible cannibalization and the minimum is taken over the communicating classes of Q_i. The minimum condition within Q_i can be interpreted as asserting that a single communicating class within Q_i is responsible for the value of $\phi T \pi_i(v)$, in the sense that the value remains unchanged if all inoperable locations in other classes suddenly become operable. As with the minimum condition, the minimum condition within Q_i is slightly stronger than this assertion.

Before presenting the result of this section, we shall introduce the notation $z_s(v) = |s| - w_s(v)$, where $|s|$ denotes the cardinality of the set s. Thus, $z_s(v)$ denotes the number of shortages in set s when the binary vector is v. Let $z_i(v) = z_{Q_i}(v)$. We shall also write $m_s(k) = |s| - n_s(k)$, and $\tilde{m}_i(k) = |Q_i| - n_i(k)$.

THEOREM 7: If a structure satisfies the minimum condition within Q_i, and if for each isolated communicating class s of Q_i, $w_s(v) = 0$ implies that $\phi(v) = 0$, then for $k \geqslant 1$

$$\tilde{m}_i(k) = \min \{m_s(k)\},$$

where the minimum is taken over the isolated communicating classes of Q_i.

PROOF: Suppose that $\tilde{m}_i(k) > \min\{m_s(k)\}$. Thus, there exists an isolated communicating class s such that $\tilde{m}_i(k) > m_s(k)$. From the assumptions of the theorem, it follows that $m_s(k) < |s|$. Hence there exists a binary vector v such that $z_s(v) = m_s(k) + 1$, and $v_j = 1$ for all locations λ_j not in s. It follows that $w_i(v) \geqslant \tilde{n}_i(k)$, and hence $\phi T\pi_i(v) \geqslant k$ for T admissible. But $\pi_i(v) = v$, hence $\phi T(v) \geqslant k$. Since $w_s(v) < n_s(k)$ and since s is isolated, it follows that $\phi T\pi_s(v) < k$. But $\pi_s(v) = v$, hence $\phi T(v) < k$. This is a contradiction, hence $\tilde{m}_i(k) \leqslant \min\{m_s(k)\}$.

Consider now any binary vector v such that $z_i(v) = \min\{m_s(k)\}$. We first wish to demonstrate that there exists a $v' \epsilon [v]$ such that $v'_j = 1$ for any location λ_j in a nonisolated communicating class of Q_i. v' can be constructed as follows. Consider in turn each isolated communicating class s of Q_i. Let S denote the union of the nonisolated communicating classes such that cannibalization from locations of s to those of S is permitted (if s is closed then S is empty). Since $z_s(v) + z_S(v) \leqslant z_i(v) \leqslant m_s(k)$, it follows that the locations of s can be cannibalized so that no shortages occur in S. The number of shortages that result in s is $z_s(v) + z_S(v) \leqslant m_s(k)$. Since each nonisolated communicating class of Q_i is contained within a set S defined above for some isolated class s, it follows that the resulting binary vector v' will have the property that $v'_j = 1$ for any location in a nonisolated communicating class of Q_i, and $z_s(v) \leqslant m_s(k)$ for each isolated communicating class s of Q_i. Thus $w_s(v) \geqslant n_s(k)$ for each communicating class s of Q_i. It follows from the definition of $n_s(k)$ that $\phi T'\pi_s(v) \geqslant k$ for each communicating class s of Q_i, where T' is an admissible cannibalization on the associated closed isolated structure. From the minimum condition within Q_i if follows that $\phi T'\pi_i(v') \geqslant k$. Since $v' \epsilon [v]$, it follows that $\pi_i(v') \epsilon [\pi_i(v)]$, and from the admissibility of T on the original structure it follows that $\phi T\pi_i(v) \geqslant \phi T\pi_i(v') \geqslant \phi T'\pi_i(v) \geqslant k$. Thus $z_i(v) \leqslant \tilde{m}_i(k)$; that is, $\min\{m_s(k)\} \leqslant \tilde{m}_i(k)$. This completes the proof.

4. STOCHASTIC MODEL

4.1 The Model

In this section the stochastic model of [3] is studied in the context of our more general restriction mappings. Because the failures of parts are chance events, the states of the locations and of the system at each point in time are random variables. The probability distributions of these random variables depend upon the failure distributions of the parts, the length of time the system has been operating, spare part inventories, the cannibalization policy employed, etc.

We shall consider a system governed by a structure function ϕ, and assume that at time $t=0$ there are $c_i (c_i \geqslant -Q_i)$ spares of part type i available for $i = 1, \ldots, N$. We interpret $c_i < 0$ to mean that initially there are $|c_i|$ shortages of part type i and no spares. If $c_i \geqslant 0$ then it is assumed that initially there are no shortages of i in the system. If at any time a failure occurs and a spare of the same type of part is available, it is assumed that a spare is installed in the failed location. After the supply of spares of a given part type has been exhausted, a failure of that type is serviced by performing the part interchanges specified by an admissible cannibalization policy. This may involve performing zero, one, or more part interchanges. For mathematical simplicity, we assume that interchanges and replacements of failed parts by spares are performed instantaneously. We also assume that failures are detected instantaneously, and that parts are never replaced until they fail.

We shall focus attention on answering the question: At a given moment of time, what is the probability that the system is in state k, for $k=0, \ldots, M$. In most cases, it is not possible to calculate these probabilities in a simple manner. If Q_i is itself a communicating class for each part type i, then an exact expression for these probabilities is given in [3]. If each Q_i is not a communicating class, but each communicating class is closed and isolated and the minimum condition within each Q_i is satis-

fied, an exact expression for these probabilities is presented in [4], though we present here a more computationally efficient estimation procedure for this case. For more general restriction mappings we obtain useful upper and lower bounds on the probabilities of interest if the admissible cannibalized structure functions satisfy the overlap condition, and the admissible cannibalized structure functions of the associated unrestricted structure satisfy the minimum condition.

Let us now introduce the notation of [3] to describe the random states of the locations and of the system. Consider the stochastic process

$$V^T(t) = (V_1^T(t), \ldots, V_n^T(t)), \qquad t \geq 0,$$

where for each fixed time t, $V_j^T(t)$ is a random variable whose possible values are zero and one. We interpret $V_j^T(t)$ as the state of the part in location λ_j at time t if T is the cannibalization policy consistently employed. For any set s of locations, let $W_s(t) = w_s(V^T(t))$, and $W_i(t) = W_{Q_i}(t)$. Thus, $W_s(t)$ represents the number of operating parts in the set of locations s at time t, and $W_i(t)$ represents the number of operating parts of type i at time t. To describe the variation in time of the random state of the system, let $\phi T(t) = \phi T(V^T(t))$. Thus, $\phi T(t)$ represents the state of the cannibalized system at time t.

As is pointed out in [3], if failure distributions are chacteristics of locations rather than of part types, then an admissible cannibalization may not correspond in practice to an advisable cannibalization. This is because the criterion of admissibility involves only the maximization of the system state at the time the interchanges are performed. But if the locations of the same part type have different failure distributions, an admissible cannibalization which consistently places parts in locations with high failure rates will lead to poorer system performance than a cannibalization that does not. We shall impose the following two restrictions to keep the mathematical analysis tractable [3].

(a) For each i, the locations of Q_i are indistinguishable in their effects on the lifetimes of the parts installed in them.

(b) The lifetime of a given part is independent of the lifetimes of any other parts.

Although assumptions (a) and (b) may in some cases represent only an approximation to the true state of affairs, the more complicated mathematical theory necessary if we do not make these assumptions will rarely be worth the effort for real systems. In practice, failure distributions are estimated from usage data which are usually kept only on a part-type basis.

Assumption (a) implies that the joint distribution of $(W_1(t), \ldots, W_N(t))$ is independent of the particular admissible cannibalization employed, and (a) and (b) together imply that the N stochastic processes

$$\{W_1(t), t \geq 0\}, \ldots, \{W_N(t), t \geq 0\} \tag{4.1}$$

are mutually independent [3].

It should be pointed out that even with assumptions (a) and (b), unless all communicating classes are closed and isolated, the distribution of $\phi T(t)$ may depend upon the particular admissible cannibalization T employed. This observation is examined in [5].

4.2 Bounds

In this section, we shall consider structures satisfying the conditions of Theorem 6. For such structures we have from Theorem 6, that for any admissible cannibalization T,

$$\sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v'|w_i(v') \leqslant \tilde{n}_i(k)\}}(v) \leqslant \phi T(v) \leqslant \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v'|w_i(v') \geqslant n_i(k)\}}.$$

Thus,

$$\sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{W_i(t) \geqslant \tilde{n}_i(k)\}} \leqslant \phi T(t) \leqslant \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{W_i(t) \geqslant n_i(k)\}}. \tag{4.2}$$

For ease of notation, we have used the convention

$$I_{\{W_i(t) \geqslant x\}} = I_{\{V^T(t)|W_i(t) \geqslant x\}}(V^T(t)).$$

The indicator functions appearing in expression (4.2) are independent of the particular admissible cannibalization T employed.

For $k=1,2, \ldots, M$ we define

$$I_k = I_{\bigcap_{i=1}^{N} \{W_i(t) \geqslant n_i(k)\}} \qquad \text{and} \qquad I'_k = I_{\bigcap_{i=1}^{N} \{W_i(t) \geqslant n_i(k)\}}.$$

Since $n_i(1) \leqslant n_i(2) \leqslant \ldots \leqslant n_i(M)$, and $\tilde{n}_i(1) \leqslant \tilde{n}_i(2) \leqslant \ldots \leqslant \tilde{n}_i(M)$, it follows that

$$I_1 \geqslant I_2 \geqslant \ldots \geqslant I_M \qquad \text{and} \qquad I'_1 \geqslant I'_2 \geqslant \ldots \geqslant I'_M. \tag{4.3}$$

From (4.2) we have

$$\left\{\sum_{k=1}^{M} I'_k \geqslant j\right\} \text{ implies that } \{\phi T(t) \geqslant j\} \text{ implies that } \left\{\sum_{k=1}^{M} I_k \geqslant j\right\}. \tag{4.4}$$

From (4.3) and the fact that $I_k = 0$ or 1 and $I'_k = 0$ or 1, it follows that

$$\left\{\sum_{k=1}^{M} I'_k \geqslant j\right\} \quad \text{if and only if} \quad \{I'_j = 1\},$$

and

$$\left\{\sum_{k=1}^{M} I_k \geqslant j\right\} \quad \text{if and only if} \quad \{I_j = 1\}.$$

Hence, (4.4) becomes

$$\{I'_j = 1\} \text{ implies that } \{\phi T(t) \geqslant j\} \text{ implies that } \{I_j = 1\}.$$

Consequently, the probability distributions of $\phi T(t)$ and $W_i(t)$ for $i = 1, \ldots, N$, satisfy the relations,

$$P\left[\bigcap_{i=1}^{N} \{W_i(t) \geqslant \tilde{n}_i(j)\}\right] \leqslant P[\phi T(t) \geqslant j] \leqslant P\left[\bigcap_{i=1}^{N} \{W_i(t) \geqslant n_i(j)\}\right].$$

Since the stochastic processes (4.1) are assumed to be mutually independent, it follows that for $j = 0, 1, \ldots, M$

$$\prod_{i=1}^{N} P[W_i(t) \geqslant \tilde{n}_i(j)] \leqslant P[\phi T(t) \geqslant j] \leqslant \prod_{i=1}^{N} P[W_i(t) \geqslant n_i(j)]. \tag{4.5}$$

Expression (4.5) is the key expression in that it gives an upper and lower bound on the probability that the state of the cannibalized system at time t is no worse than j. It is important to notice that these bounds depend only on the probability distributions of $W_i(t)$ for $i = 1, \ldots, N$.

Expressions are derived in [3] for the factors $P[W_i(t) \geqslant k]$. By assumption (a) of section 4.1, the distribution of $W_i(t)$ is independent of the cannibalization policy employed. In fact, the distribution remains unchanged if no cannibalization is employed. Hence, the expressions of [3] are applicable here. For the case $c_i \leqslant 0$, let $a = |Q_i| + c_i$. It is easily seen that for $k = 0, 1, \ldots, a$

$$P[W_i(t) = k] = \binom{a}{a-k} [Fi(t)]^{a-k} [1 - F_i(t)]^k, \tag{4.6}$$

where $F_i(t)$ denotes the lifetime distribution function for part type i. For $k > a$, $P[W_i(t) = k] = 0$.

For $c_i > 0$, results are obtained in [3] only for the case where part lifetimes follow the exponential probability law,

$$F_i(t) = 1 - e^{-z_i t} \qquad \text{for} \qquad t \geqslant 0,$$

where $1/z_i$ represents the mean time to failure for part type i. Letting $a = |Q_i|$, the result is that for $k = 1, 2, \ldots, a$

$$P[W_i(t) = a - k] = e^{-z_i a t} a^{c_i} a! / (a-k)! \sum_{j=k}^{\infty} (z_i t)^{c_i + j} S(j, k) / (c_i + j)!. \tag{4.7}$$

The $S(j, k)$ are Stirling numbers of the second kind given by

$$S(j, k) = \frac{1}{k!} \sum_{r=1}^{k} \binom{k}{r} (-1)^{k-r} r^j.$$

It is also shown in [3] that

$$Pr[W_i(t) = a] = \sum_{j=0}^{c_i} e^{-z_i a t} (z_i a t)^j / j!. \tag{4.8}$$

Hence, the bounds in expression (4.5) can be evaluated by using (4.6) for i such that $c_i \leqslant 0$ and using (4.7) and (4.8) for i such that $c_i > 0$. An approximation to (4.7) useful for large positive values of c_i and $|Q_i|$ is obtained by assuming that failures occur as in a renewal process with no decrease in total failure rate as shortages occur.

4.3 Special Cases

It was shown in [3] for the case in which complete interchangeability is permitted among the locations associated with the same part type, that if the admissible cannibalized structure function ϕT satisfies the minimum condition, then for $j = 0, 1, \ldots, M$

$$P[\phi T(t) \geqslant j] = \prod_{i=1}^{N} P[W_i(t) \geqslant n_i(j)].$$

We wish to consider here the case in which all communicating classes are closed and isolated, the admissible cannibalized structure function ϕT satisfies the minimum condition and the system satisfies the minimum condition within each Q_i. The following result is required.

THEOREM 8: If the admissible structure function ϕT satisfies the minimum condition, if the system satisfies the minimum condition within each Q_i, and if all communicating classes are closed and isolated, then

$$\phi T = \sum_{k=1}^{M} \prod_{i=1}^{N} \prod_{s \varepsilon Q_i} I_{\{v' | w_s(v') \geqslant n_s(k)\}}$$

where the final product is over the set of communicating classes within Q_i.

PROOF: It follows from Lemma 1 that

$$\phi T = \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\{v' | \phi T \pi_i(v') \geqslant k\}}.$$

It is shown in Lemma 1 of [5] that if ϕT satisfies the minimum condition and the system satisfies the minimum condition within each Q_i, then $\phi T(v) \geqslant k$ if and only if $w_s(T(v)) \geqslant n_s(k)$ for each communicating class s. It thus follows that $\phi T \pi_i(v) \geqslant k$ if and only if $w_s(T\pi_i(v')) \geqslant n_s(k)$ for all communicating classes s within Q_i. Since each s of Q_i is closed and isolated $w_s(T\pi_i(v')) = w_s(v')$. The result follows.

It follows from Theorem 8 that

$$\phi T = \sum_{k=1}^{M} \prod_{i=1}^{N} I_{\bigcap_{s \varepsilon Q_i} \{W_s(t) \geqslant n_s(k)\}}$$

where the final product is over the communicating classes of Q_i. In the same manner that (4.5) was derived from (4.2), it is easily shown that for $j = 0, 1, \ldots, M$

(4.9)
$$P\left[\phi T \geqslant j\right] = \prod_{i=1}^{N} P\left[\bigcap_{s \varepsilon Q_i} \left\{W_s(t) \geqslant n_s(j)\right\}\right].$$

Consider now the evaluation of

(4.10)
$$P[\bigcap_{s \varepsilon Q_i} \{W_s(t) \geqslant n_s(j)\}].$$

If $c_i \leqslant 0$, then (4.10) is merely the product over the communicating classes of cumulative Binomial probabilities,

(4.11)
$$\prod_{s \varepsilon Q_i} \sum_{k=0}^{b_s} \binom{a_s}{k} [F_i(t)]^k [1 - F_i(t)]^{a_s - k},$$

where $a_s = |s| + c_s$, $b_s = m_s(j) + c_s$, and $-c_s$ denotes the initial number of shortages in class s. If some $b_s < 0$, then the factor is taken as zero.

For $c_i > 0$, an exact expression for (4.10) is derived in [4], but is of a quite complex form. We shall present here an approximation that may be used when failure distributions are Exponential. Let Y_i denote the time of stockout of spares of part type i, and let F_{Y_i} denote its probability distribu-

tion function. If part lifetimes are Exponentially distributed, then expression (4.10) equals

$$\int_0^t P\left[\bigcap_{s \subset Q_i} W_s(t) \geq n_s(j) \mid Y_i = y\right] dF_{Y_i}(y) + (1 - F_{Y_i}(t)). \tag{4.12}$$

The conditional probability under the integral in (4.12) equals the product of the probabilities that in each communicating class s of Q_i there are no more than $m_s(j)$ failures from the time of stockout up till t. These probabilities are independent, and for class s equals the cumulative Binomial probability of no more than $m_s(j)$ failures out of $|s|$ operable parts each with a failure probability of $F_i(t-y)$ where $F_i(\ \ )$ is the distribution function for the Exponential lifetime distribution of part type i. Expression (4.12) thus equals

$$\int_0^t dF_{Y_i}(y) \prod_{s \subset Q_i} \sum_{d=0}^{m_s(j)} \binom{|s|}{d} [F_i(t-y)]^d [1-F_i(t-y)]^{|s|-d} + (1-F_{Y_i}(t)). \tag{4.13}$$

It is well known that $F_i(\ \ )$ Exponential with mean time to failure $1/x_i$ implies that $F_{Y_i}(\ \ )$ is the Gamma distribution function with parameters s_i and $|Q_i|/x_i$. If the communicating classes are large, then expression (4.13) can be made more computationally efficient by applying the identity [2].

$$\sum_{x=0}^{k} \binom{n}{x} p^x (1-p)^{n-x} = 1 - n\binom{n-1}{k} \int_0^p t^k (1-t)^{n-k-1}\, dt \tag{4.14}$$

and performing numerical integration to evaluate the incomplete Beta function on the right hand side of (4.14). The identity (4.14) may also be used in the evaluation of (4.11). Alternatively, the Binomial terms in (4.11) and (4.13) may be approximated by Poisson terms. This approximation is accurate when $|s|$ is large and $F_i(t)$ or $F_i(t-y)$ is small. The Poisson approximation was found to be quite useful in the study of [1] and [6].

REFERENCES

[1] Boll, C. H., and R. M. Simon, "Combining Analytic and Simulation Models to Relate Logistic Operations to Weapon System Performance," presented at the NATO Conference on Problems in the Organization and Introduction of Large Logistic Support Systems, Luxembourg (Mar. 1970).

[2] Feller, W., *An Introduction to Probability Theory and Its Applications* (John Wiley and Sons, New York, 1962), 2nd ed., p. 163.

[3] Hirsch, W. M., M. Meisner, and C. H. Boll, "Cannibalization in Multicomponent Systems and the Theory of Reliability," Nav. Res. Log. Quart. *15*, 331–360 (Sept. 1968).

[4] Simon, R. M., "The Reliability of Multi-State Systems Subject to Cannibalization," Report AM-69-1, School of Engineering and Applied Science, Washington University, St. Louis, Missouri.

[5] Simon, R. M., "Optimal Cannibalization Policies for Multicomponent Systems," Siam J. of Applied Math. *19*, 700–711 (Dec. 1970).

[6] Stanfield, H., and D. Tate, Unpublished memorandum, General Electric Center for Advanced Studies, Santa Barbara, California (1968).

BAYES ADAPTIVE CONTROL OF TWO-ECHELON INVENTORY SYSTEMS I: DEVELOPMENT FOR A SPECIAL CASE OF ONE-STATION LOWER ECHELON AND MONTE CARLO EVALUATION

S. Zacks

Case Western Reserve University

and

J. Fennell

The George Washington University

ABSTRACT

Bayes adaptive control policies are developed in the present paper for the special case of a one-station lower echelon; a Poisson distribution of demand, whose mean is assumed to have a prior gamma distribution. The cost structure is of a common type. The ordering policy for the upper echelon, which minimizes expected cost, is replaced by a new type of policy, called Bayes prediction policy. This policy does not require tedious computations, of the sort required by dynamic programming solutions. The characteristics of the policies are studied by Monte Carlo simulation, and supplemented by further theoretical development.

0. INTRODUCTION

In a previous technical memorandum on multi-echelon inventory control (see [2]) a general theory of adaptive ordering policies has been developed for cases in which the demand distribution is not completely known. In the present paper we treat a special case of a multi-echelon system which consists of one station only at the lower echelon. The monthly demand is assumed to be a random variable having a Poisson distribution with an unknown mean, λ. For a special inventory cost function, assuming no delivery costs between echelons, we develop the optimal ordering policy of the lower echelon. In the adopted Bayesian adaptive framework, the unknown parameter λ has a prescribed prior gamma distribution. The associated distribution functions are derived in Section 2, and the optimal ordering policy of the lower echelon is discussed in Section 3. In the previous paper [2] we have shown that the ordering policy of the upper echelon, which minimizes the total expected cost, can be obtained by the method of dynamic programming. This procedure requires, however, involved computations. We have therefore considered here an alternative objective for the upper echelon. This is the objective of ordering at the beginning of each month a quantity which will guarantee, with a high confidence probability, that after the order has arrived there would be sufficient number of units in stock, so that the requirement of the lower echelon could be satisfied. The lead time is two months for the upper echelon and one month for the lower echelon. The ordering policy of the upper echelon, which attains the above objective, is called a *Bayes prediction policy*. In Section 4, we develop the formulae of the Bayes prediction policy. In Section 5, we present the results of a Monte Carlo simulation, designed to provide estimates on the performance of the ordering policies under consideration. Several interesting characteristics have been observed during this simulation. As a result of the

simulation study an approximation to the expected moving monthly average costs has been developed theoretically. The theory and the basic assumptions of this approximation are presented in Section 6. We provide also numerical computations of the approximating functions. Section 7 is devoted to test the sensitivity of the Bayes prediction policy. Expected moving monthly averages were estimated, on the basis of the simulation results, when the Bayes prediction ordering values were systematically increased or decreased. The results of these sensitivity tests reinforce our general conclusion that the Bayes prediction policy, as an alternative to the dynamic programming policy, is an effective control policy. The results of the present study indicate that our adaptive procedures keep the inventory system very stable, with only minor fluctuations, even if the demand distributions are unknown or sometimes unstationary.

1. THE INVENTORY SYSTEM

The inventory system under consideration consists of two echelons. An upper echelon, D, at which shipments from manufacturers arrive, and a lower echelon, E, at which customers are supplied. The stock levels at D and at E are adjusted at the beginning of each month, at which new orders are placed. Let S_n designate the stock level at the whole system at the beginning of the nth month, and Q_n the corresponding stock level at the lower echelon, E. Let $X_1, X_2, \ldots$ designate a sequence of independent and identically distributed random variables, which present the monthly demand (for the item under consideration). Further assumptions on the distribution of X_n will be specified in Section 2. The lower echelon can order units from the upper echelon, or can send back undemanded stock. Thus, let Y_n designate the quantity ordered by the lower echelon, from the upper echelon, at the beginning of the nth month. Y_n may be negative, in cases of back shipping of stock, and satisfies the inequality

$$-Q_n \leq Y_n \leq S_n - Q_n, \qquad n = 1, 2, \ldots. \tag{1.1}$$

The quantity ordered by E at the beginning of a month arrives at its destination at the *end* of that month. Let Z_n designate the quantity ordered by the upper echelon, D, from the manufacturers, at the beginning of the nth month. We assume that the lead time for such orders is 2 months. That is, the quantity Z_n arrives at D at the end of the $(n+1)$st month. In the present paper we consider an inventory cost function of the same type as in [2], but somewhat simpler. The cost associated with the ordering policy is presented by the simple cost function

$$C(Z_n, X_n, Q_n) = c^* Z_n + \Phi(Q_n - X_n), \tag{1.2}$$

where C^* is the cost [$] of ordering Z_n units, and $\Phi(\cdot)$ is a convex function representing the cost of shortage or of surplus of stock. The effect of Y_n on the accrued cost is via the stock level Q_n. The exact relationship will be discussed later.

2. DISTRIBUTIONS EMPLOYED IN THE DEVELOPMENT OF THE BAYESIAN ADAPTIVE CONTROL POLICIES

We assume here that the random variables $X_1, X_2, \ldots$, which represent the monthly demand, are independent and have an identical Poisson distribution with an unknown mean λ, $0 < \lambda < \infty$.

After observing the demand for n months, we consider the minimal sufficient statistic, T_n, which is the sum

$$T_n = \sum_{i=1}^{n} X_i, \qquad n=1,2,\ldots; \qquad \text{and} \qquad T_0 \equiv 0.$$

We assume in our Bayesian model that the expected monthly demand, λ, has a prior gamma distribution $G\left(\frac{1}{\tau}, \nu\right)$; $0<\tau, \nu<\infty$, whose density function is

$$h(\lambda|\tau,\nu) = \frac{1}{\Gamma(\nu)\tau^{\nu}} \lambda^{\nu-1} \exp\{-\lambda/\tau\}, \; 0<\lambda<\infty. \tag{2.1}$$

As shown previously [2], the anticipated distribution of the random variable X_{n+1}, given T_n, is the mixture of Poisson distributions with respect to the posterior distribution of λ given T_n, and is the negative-binomial distribution, with a probability density

$$g(x|\psi_{n+1},\nu_{n+1}) = \frac{\Gamma(\nu_{n+1}+x)}{\Gamma(x+1)\Gamma(\nu_{n+1})} (1-\psi_{n+1})^{\nu_{n+1}} \psi_{n+1}^{x}, \qquad x=0,1,\ldots, \tag{2.2}$$

where

$$\psi_{n+1} = \frac{\tau}{1+(n+1)\tau}, \qquad n=0,1,\ldots \tag{2.3}$$

$$\nu_{n+1} = \nu + T_n, \qquad n=0,1,\ldots.$$

We shall designate a negative-binomial distribution with parameters ψ and ν; $0<\psi<1, 0<\nu<\infty$, by N.B. (ψ, ν). Since, for a given λ, the random variables $X_{n+1}, \ldots X_{n+k} (k \geq 2)$ are independent, it is easy to verify that the anticipated distribution of $X_{n+1} + \ldots + X_{n+k}$, $k \geq 2$, given T_n is the *N.B.* $(k\psi_{n+k}, \nu+T_n)$; where as in (2.3), $\psi_{n+k} = \tau/[1+(n+k)\tau]$. Another anticipated distribution playing an important role is that of (X_{n+1}, X_{n+2}) given T_n. Since the anticipated distribution of X_{n+2} given (T_n, X_{n+1}) is the *N.B.* $(\psi_{n+2}, \nu+T_n+X_{n+1})$ and that of X_{n+1} given T_n is *N.B.* $(\psi_{n+1}, \nu+T_n)$, the joint probability density of (X_{n+1}, X_{n+2}) given T_n is

$$g(x,y|\psi_{n+1}, v+T_n) = g(x|\psi_{n+1}, v+T_n) g(y|\psi_{n+2}, v+T_n+x); \tag{2.4}$$

for all $x, y = 0,1,\ldots$.

3. EXPLICIT DETERMINATION OF THE OPTIMAL ORDERING POLICY FOR THE LOWER ECHELON

As shown in [4], the optimal ordering level for the lower echelon, Y_n^0, at the beginning of the nth month, does not depend on that of the upper echelon. It depends only on the state variables (Q_n, S_n, T_{n-1}); where Q_n is the stock level of the lower echelon, at the beginning of the nth month; S_n is that of the whole system; and T_{n-1} is the total demand during the previous $n-1$ months. These state variables satisfy the following recursive relations:

$$Q_{n+1} = (Q_n + Y_n^0 - X_n)^+, \qquad n=1,2,\ldots,$$

$$S_{n+1} = (S_n + Z_{n-1}^* - X_n)^+, \qquad n=1,2,\ldots,$$

and

$$T_n = T_{n-1} + X_n, \qquad n = 1, 2, \ldots,$$

where $T_0 \equiv 0$ and Z^*_{n-1} is the quantity ordered by the upper echelon at the beginning of the $(n-1)$st month. If no delivery cost is charged to the lower echelon then, as shown in [2], the optimal ordering level of the lower echelon at the beginning of the nth month is

$$Y^o_n(T_{n-1}, S_n, Q_n) = K_n(T_{n-1}, S_n) - Q_n, \qquad n = 1, 2, \ldots, \text{ where} \tag{3.1}$$

$$K_n(T_{n-1}, S_n) = \min\{S_n, k^o_n(T_{n-1})\}, \text{ and} \tag{3.2}$$

$$k^o_n(T_{n-1}) = \text{least nonnegative integer}, k, \text{ such that} \tag{3.3}$$

$$\Delta_k E^{T_{n-1}}\{\Phi((k - X_n)^+ - X_{n+1})\} \geq 0.$$

The difference function $\Delta_k F(k)$ is defined as usual as $\Delta_k F(k) = F(k+1) - F(k)$, $k = 0, 1, \ldots$; and as before $\Phi(\cdot)$ is a convex inventory cost function associated with the lower echelon. $E^{T_{n-1}}\{\cdot\}$ designates the conditional expectation, given T_{n-1}. In the present paper, we consider the cost function $\Phi(t) = ct^+ + pt^-$, where $t^+ = \max(t, 0)$ and $t^- = -\min(t, 0)$; $0 < c, p < \infty$. Let $I\{A\}$ designate the indicator function of the set A. It is not difficult to verify that

$$\Delta_k \Phi((k - X_n)^+ - X_{n+1}) = cI\{k \geq X_n + X_{n+1}\} - pI\{X_n \leq k < X_n + X_{n+1}\}; \qquad k = 0, 1, \ldots. \tag{3.4}$$

Hence,

$$\Delta_k E^{T_{n-1}}\{\Phi((k - X_n)^+ - X_{n+1})\} = cP[k \geq X_n + X_{n+1} | T_{n-1}] - pP[X_n \leq k < X_n + X_{n+1} | T_{n-1}]. \tag{3.5}$$

Let $G(x|\psi, \nu)$ designate the *c.d.f.* of *N.B.* (ψ, ν) at the point x. Since $P[X_n \leq k < X_n + X_{n+1} | T_{n-1}] = P[X_n \leq k | T_{n-1}] - P[X_n + X_{n+1} \leq k | T_{n-1}]$ we obtain, by employing the results mentioned in the previous section, that

$$k^o_n(T_{n-1}) = \text{least non-negative integer}, k, \text{ such that} \tag{3.6}$$

$$G(k | 2\psi_{n+1}, \nu + T_{n-1}) \geq \frac{p}{c+p} G(k | \psi_n, \nu + T_{n-1}).$$

The behavior of the function $k^o_n(T_{n-1})$, $n = 1, 2, \ldots$ will be discussed and illustrated numerically in Section 5. A FORTRAN program for the computation of $k^o(T_n)$ is described in [1].

4. AN ORDERING POLICY FOR THE UPPER ECHELON, BASED ON BAYESIAN PREDICTION

In Section 4 of [2], we developed the dynamic programming equations, which are required for the determination of the optimal ordering levels of the upper echelon. These dynamic programming equations require a tabulation of minimal Bayes posterior expected cost as functions of two-state variables. In models where the lead-time is larger than 2 months the number of state variables is larger than 2.

For these reasons, we approached the problem of determining an ordering policy for the upper echelon from a different point of view. The policy which we develop is not a myopic Bayes policy, in the sense that we reiterate the policy which minimizes the posterior expected inventory cost of the last ordering period, but a reiteration of a different policy. For this reason, we have called our policy a Bayes prediction policy. The following is the idea: We have seen in (3.1) that the order level of the lower echelon at the beginning of the $(n+2)$nd month is

$$(4.1) \quad K_{n+2}(((S_n^*-X_n)^+ + Z_n - X_{n+1})^+, T_{n-1}+X_n+X_{n+1}) = \min\{k_{n+2}^o(T_{n-1}+X_n+X_{n+1}), ((S^*-X_n)^+ + Z_n - X_{n+1})^+\},$$

where $S_n^* = S_n + Z_{n-1}$; Z_{n-1} is the order of the upper echelon at the beginning of the $(n-1)$st month; and Z_n is the corresponding order level at time n. Thus, the quantity Z_n ordered by the upper echelon, at the beginning of the nth month, will have its effect on the optimal ordering level of the lower echelon at the beginning of the $(n+2)$nd month. It is desirable that the stock level of the whole system at the beginning of the $(n+2)$nd month will yield the equality

$$(4.2) \quad K_{n+2}((S_n^* - X_n)^+ + Z_n - X_{n+1})^+, T_{n-1}+X_n+X_{n+1}) = k_{n+2}^o(T_{n-1}+X_n+X_{n+1}).$$

No policy $Z_n = Z(S_n^*, T_{n-1})$ can guarantee (4.2) with certainty. However, it seems reasonable to impose the condition that

$$(4.3) \quad Z_n = Z(S_n^*, T_{n-1}) = \text{least nonnegative integer}, Z, \text{ such that}$$
$$P[Z + (S_n^* - X_n)^+ - X_{n+1} \geq k_{n+2}^o(T_{n-1}+X_n+X_{n+1}) \mid T_{n-1}] \geq \gamma;$$

for a specified prediction confidence level γ, $0<\gamma<1$. We have simplified the determination of $Z(S_n^*, T_{n-1})$ by making the following observations. As n grows, the sequence $\{k_{n+2}^o(T_{n-1}+X_n+X_{n+1}) : n \geq 1\}$ approaches a limit $k^o(\lambda)$, where

$$(4.4) \quad k^o(\lambda) = \text{least non-negative integer}, k, \text{ such that } P(k; 2\lambda) \geq \frac{p}{c+p} P(k; \lambda);$$

$P(k; \lambda)$ is the *c.d.f.* of a Poisson distribution with mean λ, and $P(k; 2\lambda)$ is that of a Poisson distribution with mean 2λ. This is implied by the fact that, as $n \to \infty$, the anticipated distribution of X_n, and of $X_n + X_{n+1}$, given T_{n-1} converge strongly to the true Poisson distribution of the observed demand. As will be illustrated later numerically, although it is certain that $k_{n+2}^o(T_{n-1}+X_n+X_{n+1}) \geq k_n^o(T_{n-1})$ for all $n = 1, 2, \ldots$ the difference between these functions tends to zero as n grows. We therefore simplified our procedure by substituting $k_n^o(T_{n-1})$ instead of $k_{n+2}^o(T_{n-1}+X_n+X_{n+1})$ in formula (4.3). Thus, we consider the following ordering policy for the upper echelon,

$$(4.5) \quad Z_n^* = \text{least nonnegative integer}, Z, \text{ such that } P[Z + (S_n^* - X_n)^+ - X_{n+1} \geq k_n^o(T_{n-1}) \mid T_{n-1}] \geq \gamma; \quad n = 1, 2, \ldots .$$

As shown in [3], the conditional probability term of (4.5) is given by

$$P[Z+(S_n^*-X_n)^+-X_{n+1} \geq k_n^o(T_{n-1}) \mid T_{n-1}] = G(Z-k_n^o(T_{n-1}) \mid \psi_n, \nu+T_{n-1}) \tag{4.6}$$

$$+ I\{S_n^*>0\} \sum_{x=0}^{S_n^*-1} g(x \mid \psi_n, \nu+T_{n-1})[G(Z+S_n^*-x-k_n^o(T_{n-1}) \mid \psi_{n+1}, \nu+T_{n-1}+x)$$

$$-G(Z-k_n^o(T_{n-1}) \mid \psi_{n+1}, \nu+T_{n-1}+x)] .$$

The ordering policy Z_n^* is called a *Bayes prediction ordering policy.* Its characteristics will be discussed later. For the numerical determination of Z_n^* consult [1].

5. MONTE CARLO EVALUATION OF THE ORDERING POLICIES

In order to investigate the actual behavior, and to estimate the total expected inventory cost of a two-echelon system, which is subjected to the ordering policies developed above, we have simulated four cases:

Case I. $\lambda=10$, $c=0.5$, $p=10$;
Case II. $\lambda=10$, $c=10$, $p=10$;
Case III. $\lambda=1$, $c=0.5$, $p=10$;
Case IV. $\lambda=1$, $c=10$, $p=10$.

In each of the four cases we have made 10 independent runs, each run consisting of 50 months. Although it might seem at the outset that 10 independent runs cannot furnish conclusive evidence, the intensive analysis to which we have subjected the simulation results indicate that sufficient information is available. Computational details are provided in Fennell's paper [1]. In our simulation we have initiated the controls at the beginning of the 11th month. The reason for starting the actual control at the 11th month, rather than at the 1st month, is to enable the investigation of the rectifying effect of our policies on systems which, for a certain period of time, have not been subjected to the type of control that our ordering policies provide. In each one of the 40 runs of our Monte Carlo experiment, the stock levels, at the beginning of the 11th month, are $S_0=10$ and $Q_0=5$ (units). The total demand, T_{10}, during the first 10 months differs, however, at random from one run to another. In our presentation of the results we will consider the 11th month as the first control month, and present the behavior characteristics over a period of 25 months from the initial control month. We could interpret the simulation results also as though they were obtained from systems with the same initial stock levels (S_0, Q_0), but with different values of the prior parameter $\nu_0=\nu+T_0$. Independent and identically distributed random variables $X_1, X_2, \ldots$ having Poisson distributions with specified means have been simulated according to a procedure described in [1]. On the basis of the generated X-values we computed stock levels S_n and Q_n; ordering levels Y_n^o and Z_n^*; monthly costs C_n and moving averages of monthly costs $E_n=\sum_{i=1}^{n} C_i/n$. In Table 1 we present the output of a typical single simulation run.

Table 1 indicates certain interesting characteristics, which will be discussed briefly. The important conclusions are underlined.

(i) *The function* $k_n^o(T_{n-1})$, which determines the desired stock level at the lower echelon, *varies slowly with* T_{n-1} *and* n. Almost all the $k_n^o(T_{n-1})$ values in Table 1 are equal to 25. The slow variation of $k_n^o(T_{n-1})$, which has been consistently observed in the various runs, supports the supposition made in Section 4 that, in many instances, $k_{n+2}^o(T_{n-1}+X_n+X_{n+1})=k_n^o(T_{n-1})$.

TABLE 1. *Simulated Inventory Statistics Case I*, $\nu_0 = 97.5$, $\tau = 11$, $\gamma = 0.85$

n	X_n	k_n^o	S_n	Q_n	Y_n^o	Z_n^*	C_n	E_n
1	5	26	10	5	5	37	259.00	259.00
2	10	25	5	5	0	5	85.00	172.00
3	8	25	32	0	25	10	150.00	164.66
4	10	25	29	17	8	8	59.50	138.37
5	6	25	29	15	10	10	74.50	125.60
6	5	25	31	19	6	6	49.00	112.83
7	8	24	36	20	4	3	27.00	100.57
8	12	24	34	16	8	8	58.00	95.25
9	9	25	25	12	13	13	92.50	94.94
10	11	25	24	16	8	10	72.50	92.70
11	9	25	26	13	12	11	79.00	91.45
12	6	25	27	16	9	9	68.00	89.50
13	9	25	32	19	6	6	47.00	86.23
14	9	25	32	16	9	9	66.50	84.82
15	11	25	29	16	9	9	65.50	83.53
16	13	25	27	14	11	11	77.50	83.15
17	4	25	23	12	11	13	95.00	83.85
18	11	25	30	19	6	4	32.00	80.97
19	8	25	32	14	11	11	80.00	80.92
20	13	25	28	17	8	8	58.00	79.78
21	14	25	26	12	13	13	111.00	81.26
22	10	26	20	11	9	16	112.50	82.68
23	7	26	23	10	13	10	71.50	85.93
24	21	26	32	16	10	6	92.00	86.20
25	11	26	21	5	16	22	214.00	87.86

(ii) With a known value of λ one should compute the optimal stock level at the lower echelon $k^o(\lambda)$, according to formula (4.4). In the present case of $\lambda = 10$ we have $k^o(10) = 28$. In Table 1, however, we observe $k_n^o(T_{n-1})$ values which are almost consistently equal to 25. This indicates that our control procedure keeps the stock level at the lower echelon in line with the history of the demand. In the specific run illustrated in Table 1, the demand has been consistently smaller than anticipated. This has been immediately reflected by lower values of $k_n^o(T_{n-1})$. *It seems that the adaptive ordering policy considered here may provide a good protection against deviations in the demand distribution from the supposed law.* The problem of evaluating the robustness of the adaptive ordering policy under consideration will be studied in another paper.

(iii) *The stock levels* S_n *and* Q_n *are adjusted rapidly to the appropriate levels, and stay on that level with only minor fluctuations.* The adjustment of S_n and Q_n is provided by an immediate reaction of the Z_n^* function to excessively low or high stock levels. For example, the initial stock levels in our simulation runs were $S_0 = 10$ and $Q_0 = 5$. In the cases of $\lambda = 10$ the stock levels should be higher. The function $Z_1^*(T_o)$ adjusts these low initial stock levels immediately. Due to the lead-time of 2 months

TABLE 2. *Sample Statistics of* $Z_1^*(T_o)$

Case	Minimum	Median	Maximum
I	32	37	42
II	24	29	34
III	0	0	1
IV	0	0	0

between placing the orders of the upper echelon until they arrive at the system, the values of $Z_1^*(T_o)$ are above those of S_n and Q_n. In the cases of $\lambda=1$ the initial stocks are too high. This fact is reflected by small values of $Z_1^*(T_o)$. In Table 2 we provide some statistics which represent the distributions of $Z_1^*(T_o)$ for the various cases.

(iv) We have observed in Table 1 that once the first adjustment of stock levels has been attained, *the values of* $Z_n^*(T_{n-1})$ *are very close, in many cases even equal, to those of* X_{n-1}. In Table 3, we present the joint frequency distributions of the pairs (X_{n-1}, Z_n^*), $n=2(1)30$, over the 10 replicas of Case I. Similar tables for Cases II–IV can be found in [3]. We see in the table that the values of Z_n^* are highly dependent, in the statistical sense, on the values of X_{n-1}. The dependence is nearly linear. The sample statistics of the X-values and those of the Z_n^* values for the four cases under consideration are:

Statistics	Case I	Case II	Case III	Case IV
$\bar{X}$	9.00	9.00	1.00	1.00
σ_X	3.13	3.12	1.03	1.03
$\bar{Z}$	9.00	9.00	0.00	0.00
σ_Z	3.73	3.75	1.15	0.98
ρ_{XZ}	0.974	0.982	0.935	0.794

TABLE 3. *Joint Frequency Distribution of* (X_{n-1}, Z_n^*): $n=2(1)30$, *Case I*

$Z \backslash X$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	Σ
1	1																			1
2	0	4																		4
3	3	1	5																	9
4		2	2	4																8
5			5	5	2															12
6				10	7															17
7				1	15	11														27
8					1	24	3													28
9							35													35
10			1				1	28												30
11								3	22					1						26
12								4	7	10										21
13								1	3	11	13									28
14									1	1	5	5								12
15											4	6	4							14
16												2	3	1						6
17													2	1						3
18														1						1
19															4					4
20																2				2
21																				0
22																			2	2
Σ	4	7	13	20	25	35	39	36	33	22	22	13	9	4	4	2	0	0	2	290

$\bar{X}$, $\bar{Z}$ and σ_X, σ_Z designate the marginal means and standard deviations of the X_{n-1}'s and the Z_n^*'s in these tables, respectively. ρ_{XZ} designates the coefficient of correlation between X_{n-1} and Z_n^*. *These statistics*, each one based on samples of 290 pairs, *indicate that one can replace the ordering policy* Z_n^* *by the simplest ordering policy:* $Z_n = X_{n-1}$; namely, the order of the upper echelon can be set equal

to the demand in the previous month, *without increasing significantly the expected inventory cost.* This point will be discussed and illustrated numerically in Section 6.

(v) The moving-averages of monthly costs, E_n, $n=1,2,\ldots$, were computed in each of the simulation runs. In Table 4, we present the means of E_n, over the 10 independent runs. These means are (unbiased) estimates of the expected values of E_n. The exact expectations of E_n can be computed by elaborate recursive formulae, which we do not present in the present paper. A theoretical approximation to the expectation of E_n is developed in Section 6, and presented numerically in Table 4, too.

TABLE 4. *Simulation Means of Monthly Moving Averages,* $\bar{A}_n$, *and Theoretical Approximation* A_n^*

(The values in brackets are the standard errors of $\bar{A}_n$)

n	Case I		Case II		Case III		Case IV	
	$\bar{A}_n$	A_n^*	$\bar{A}_n$	A_n^*	$\bar{A}_n$	A_n^*	$\bar{A}_n$	A_n^*
1	345.50 (12.31)	309.45	287.40 (12.09)	253.86	3.45 (1.52)	2.01	38.00 (5.33)	40.01
2	244.40 (11.62)	235.78	215.70 (11.53)	208.00	4.87 (2.18)	5.38	25.50 (2.93)	28.51
3	208.13 (8.41)	209.67	189.23 (8.16)	191.15	5.23 (2.03)	6.49	22.47 (1.62)	24.35
4	173.89 (6.29)	176.13	173.07 (6.84)	170.00	6.41 (1.96)	7.06	21.37 (1.79)	22.27
5	155.24 (5.45)	156.43	163.70 (6.11)	158.10	7.67 (1.55)	7.41	20.72 (1.50)	21.02
6	144.59 (5.55)	143.68	160.75 (6.64)	150.60	7.85 (1.48)	7.67	19.90 (1.21)	20.19
7	136.79 (6.78)	134.87	154.71 (7.88)	145.51	8.54 (1.37)	7.86	19.71 (1.24)	19.60
8	131.57 (6.32)	128.49	153.93 (7.11)	141.86	8.64 (1.42)	8.03	19.33 (1.37)	19.15
9	124.94 (4.99)	123.71	149.34 (5.76)	139.13	8.69 (1.38)	8.17	19.29 (1.35)	18.81
10	122.16 (4.16)	120.02	149.68 (4.70)	137.03	9.02 (1.38)	8.29	19.36 (1.59)	18.53
11	119.05 (3.82)	117.12	149.47 (4.23)	135.38	9.00 (1.45)	8.40	19.56 (1.54)	18.31
12	115.99 (3.66)	114.79	147.55 (4.47)	134.04	9.17 (1.38)	8.50	19.58 (1.44)	18.12
13	112.07 (3.18)	112.89	143.53 (4.38)	132.95	9.06 (1.12)	8.59	19.43 (1.33)	17.97
14	109.11 (3.35)	111.33	141.75 (5.07)	132.04	8.75 (1.02)	8.67	18.95 (1.25)	17.83
15	106.73 (3.40)	110.03	140.11 (4.87)	131.27	8.95 (0.85)	8.74	19.00 (1.13)	17.71
16	103.70 (2.96)	108.93	137.04 (3.95)	130.61	9.04 (0.80)	8.81	18.89 (0.95)	17.62
17	102.06 (2.91)	108.01	135.44 (4.08)	130.03	9.35 (0.78)	8.87	18.92 (0.96)	17.53
18	101.48 (3.19)	107.21	135.77 (4.07)	129.54	9.17 (0.78)	8.93	18.75 (1.02)	17.45
19	101.67 (3.15)	106.52	136.08 (4.07)	129.10	9.12 (0.79)	8.98	18.82 (0.97)	17.38
20	100.44 (3.05)	105.93	135.55 (3.83)	128.70	9.01 (0.70)	9.03	18.84 (0.92)	17.31
21	100.58 (2.73)	105.41	136.04 (3.23)	128.36	9.19 (0.63)	9.07	19.00 (0.87)	17.26
22	100.82 (2.48)	104.96	136.67 (3.06)	128.04	9.20 (0.59)	9.11	19.02 (0.82)	17.20
23	100.66 (2.63)	104.56	136.57 (3.46)	127.75	9.27 (0.56)	9.15	18.91 (0.80)	17.16
24	101.77 (3.01)	104.22	137.19 (3.72)	127.49	9.45 (0.62)	9.19	18.90 (0.80)	17.11
25	102.88 (3.75)	103.90	138.00 (4.24)	127.25	9.56 (0.63)	9.22	18.97 (0.81)	17.07

6. APPROXIMATE FORMULA FOR THE EXPECTED MONTHLY MOVING AVERAGES

The expected monthly moving average, over a period of N months, is

(6.1) $$A_N = \frac{1}{N}\sum_{n=1}^{N} E\{c^* Z_n^* + c(Q_n - X_n)^+ + p(Q_n - X_n)^-\};$$

where c^* is the cost of ordering one unit.

In the previous section we have indicated that, for $n \geq 2$, Z_n^* is closely related to X_{n-1}. Therefore, in the approximation to A_N developed here, we approximate $E\{Z_n^*\}$ by $\lambda = E\{X_{n-1}\}$. For $n=1$, we have

(6.2) $$E_\lambda\{Z_1^*(T_o)\} = e^{-\lambda^*}\sum_{t=0}^{\infty}\frac{(\lambda^*)^t}{t!}Z_1^*(t),$$

where λ^* is the mean of the Poisson distribution of T_o (in the simulation cases of Section 5, $\lambda^* = 10\lambda$). The function $Z_1^*(t)$ is computed according to (4.5)–(4.8), with $S_1^* = S_o$ (the initial stock level). Theoretically, the expected value of $Z_1^*(T_o)$ is not equal to the value of $Z_1^*(t)$ at $t = E_\lambda^* \{T_o\}$. However, as we have observed from the simulation results, $E_\lambda\{Z_1^*(T_o)\}$ can be approximated very well by $Z_1([E_\lambda\{T_o\}])$, where $[\cdot]$ designates the integral part of the variable. We exhibit this in Table 5.

TABLE 5. *Approximations to* $E_\lambda\{Z_1^*(T_o)\}$

Case	$Z_1^*([\lambda^*])$	$\bar{Z}_1^*(T_o)$	S.E. $\{\bar{Z}_1^*\}$
I	37	37.5	0.934
II	29	29.2	0.934
III	0	0.1	0.010
IV	0	0.0	†0.000

†All ten $Z_1^*(T_o)$ in Case IV are equal to zero.

Therefore, in our approximation to the expected values of $A_N (N=1, 2, \ldots)$ we have used $Z_1^*([\lambda^*])$ as an approximation for $E_\lambda\{Z_1^*(T_o)\}$.

We consider now the problem of approximating $E_\lambda\{c(Q_n - X_n)^+ + p(Q_n - X_n)^-\}$, $n = 1, 2, \ldots$. According to the recursive relations of the inventory system, $Q_n = (K_{n-1} - X_{n-1})^+$, where $K_{n-1} = \min\{K_{n-1}^o(T_{n-2}), S_{n-1}\}$. Thus, Q_n is a function only of the first $n-1$ observations, and K_{n-1} is a function of the first $n-2$ observations. Thus, for each $n \geq 2$ we start by determining the conditional expectation given the first $(n-2)$ observations, which is

(6.3)

$$M_\lambda(K_{n-1}) = E_\lambda\{c(Q_n - X_n)^+ + p(Q_n - X_n)^- \mid X_1, \ldots, X_{n-2}\} = \lambda p + (c+p) I\{K_{n-1} \geq 1\} \Big[e^{-\lambda} \sum_{j=0}^{K_{n-1}-1} \frac{\lambda^j}{j!} (K_{n-1} - j) P(K_{n-1} - j; \lambda) - \lambda e^{-\lambda} \sum_{j=0}^{K_{n-1}-1} \frac{\lambda^j}{j!} P(K_{n-1} - j - 1; \lambda) - \frac{p}{c+p} [P(K_{n-1} - 1; \lambda) \cdot K_{n-1} - \lambda P(K_{n-1} - 2; \lambda)] \Big].$$

Here $I\{K_{n-1} \geq 1\}$ is the indicator function of the set $\{K_{n-1} \geq 1\}$. By definition, $P(k;\lambda) \equiv 0$ for $k < 0$. In the case of $n = 1$ we have $Q_1 = Q_0$, which is the initial stock level at the lower echelon, and

$$E\{c(Q_1 - X_1)^+ + p(Q_1 - X_1)^-\} = (c+p)[Q_1 P(Q_1;\lambda) - \lambda P(Q_1 - 1;\lambda)] + p(\lambda - Q_1). \tag{6.4}$$

As previously indicated $K_{n-1} = \min\{S_{n-1}, K_{n-1}^o(T_{n-2})\}$. We have also seen that the function $k_{n-1}^o(T_{n-2})$ varies very slowly. For this reason we simplify the approximation by determing the conditional distribution of K_{n-1}, for a fixed value of $k_{n-1}^o(T_{n-2})$. One can consider, for example, the value of $k^o \equiv k_{15}^o([E\{T_{14}\}])$. The specific values of k^o used in our computations will be given later. For $n = 2$ we approximated the expected value of $M_\lambda(K_1)$ by $M_\lambda(k_1^*)$, where $k_1^* = \min\{S_o, k^o\}$. For values of $n \geq 3$, we have derived first the distribution of S_n, and then computed $E_\lambda\{M_\lambda(K_{n-1}^*)\}$, where $K_{n-1}^* = \min\{S_{n-1}, k^o\}$.

The distribution of S_n can be determined according to the following recursive formulae:

$$
\begin{aligned}
&S_1 = S_o \\
&S_2 = (S_1 - X_1)^+ \\
&S_3 = (S_2 + Z_1^* - X_2)^+ \\
&S_n = (S_{n-1} + X_{n-3} - X_{n-1})^+, \qquad \text{all } n \geq 4,
\end{aligned} \tag{6.5}
$$

where $X_1, X_2, \ldots$ is a sequence of independent random variables having an identical Poisson distribution, with mean λ. As we have previously seen, Z_{n-2}^* is distributed very closely to X_{n-3}, for all $n \geq 4$. For this reason X_{n-3} is substituted for Z_{n-2}^*, in formula (6.5), for the cases of $n \geq 4$.

Let $h_n(x|S_{n-1})$, $n \geq 4$, designate the conditional probability distribution function of S_n, given S_{n-1}. Let $h_3(x|S_2, Z_1^*)$ designate the conditional probability distribution function of S_3, given (S_2, Z_1^*) and let $\zeta_n(x|S_o, Z_1^*)$ designate the conditional probability distribution of S_n, for given values of S_o and Z_1^*. We have

$$\zeta_1(x|S_o, Z_1^*) = I\{x = S_o\}; \tag{6.6}$$

$$\zeta_2(x|S_o, Z_1^*) = \begin{cases} 1 - P(S_o - 1; \lambda), & \text{if } x = 0 \\ p(S_o - X; \lambda), & \text{if } x \geq 1, \end{cases} \tag{6.7}$$

where $p(j;\lambda)$ is the Poisson probability distribution function with mean λ; and $p(j;\lambda) = 0$ for all $j < 0$. For $n = 3$ we have,

$$h_3(x|S_2, Z_1^*) = \begin{cases} 1 - P(S_2 + Z_1^* - 1; \lambda), & \text{if } x = 0 \\ p(S_2 + Z_1^* - x; \lambda), & \text{if } x \geq 1. \end{cases} \tag{6.8}$$

Hence,

$$\zeta_3(x|S_o, Z_1^*) = \sum_{y=0}^{S_o} \zeta_2(y|S_o, Z_1^*) h_3(x|y, Z_1^*), \qquad x = 0, 1, \ldots. \tag{6.9}$$

For $n \geq 4$ we obtain that

$$h_n(x|S_{n-1}) = \begin{cases} 1 - \sum_{y=0}^{\infty} p(y;\lambda) P(S_{n-1} + y - 1; \lambda), & x = 0 \\ f(S_{n-1} - x; \lambda), & x \geq 1, \end{cases} \tag{6.10}$$

where $f(k;\lambda)$ is the probability distribution function of $X_1 - X_2$, given by

$$f(k;\lambda) = e^{-2\lambda} \sum_{i=0}^{\infty} \frac{\lambda^{2i+k}}{i!k!}, \qquad k = 0, 1, \ldots \tag{6.11}$$

and $f(-k;\lambda) = f(k;\lambda)$ for all $k \geq 0$. (The infinite sum in (6.11) is known as a modified Bessel function of order $k, I_k(2\lambda)$). Thus, we determine $\zeta_n(x|S_o, Z_1^*)$ for $n \geq 4$, according to (6.7)–(6.10) and the recursive formula:

$$\zeta_n(x|S_o, Z_1^*) = \sum_{y=0}^{\infty} \zeta_{n-1}(y|S_o, Z_1^*) h_n(x|y); \qquad x = 0, 1, \ldots. \tag{6.12}$$

Finally, since

$$M_\lambda(K_{n-1}) = \begin{cases} M_\lambda(S_{n-1}), & \text{if } S_{n-1} \leq k_{n-2}^0(T_{n-2}) \\ M_\lambda(k_{n-1}^0(T_{n-2})), & \text{if } S_{n-1} > k_{n-1}^0(T_{n-2}) \end{cases} \tag{6.13}$$

we approximate the conditional expectation of $M_\lambda(K_{n-1})$, given Z_1^*, by

$$E_n^*(k^o, S_o, Z_1^*) = M_\lambda(k^o) + \sum_{x=0}^{k^o-1} \zeta_{n-1}(x|S_o, Z_1^*)[M_\lambda(x) - M_\lambda(k^o)]. \tag{6.14}$$

The expectation of $M_\lambda(K_{n-1})$ has been approximated by $E_n^*(k^o, S_o, Z_1^*([\lambda^*]))$, where the $Z_1^*([\lambda^*])$ values are given in Table 5. The approximation derived for the expected monthly moving average A_N, $N=1, 2, \ldots$, is therefore given by:

$$A_N^* = \frac{1}{N}\Big\{c^*Z_1^*([\lambda^*]) + (c+p)[Q_1P(Q_1;\lambda) - \lambda P(Q_1-1;\lambda)] + p(\lambda - Q_1) + (N-1)c^*\lambda + I\{N \geq 2\} \sum_{n=2}^{N} E_n^*(k^o, S_o, Z_1^*([\lambda^*]))\Big\}, \quad N=1, 2, \ldots. \tag{6.15}$$

In Table 4 we present the approximation A_N^*, to the expected moving monthly average costs A_N, for $N=1, \ldots, 25$. The parameters used for the four cases are:

Parameters	Case I	Case II	Case III	Case IV
k^o	26	18	5	1
$Z_1^*([\lambda^*])$	37	29	0	0

As shown above, all the simulated means, $\bar{A}_n$, for $n \geq 2$, in Cases I and III, are not significantly different from the theoretical approximation A_n^*. (A difference between $\bar{A}_n$ and A_n^* is considered to be significant if its magnitude is greater than two standard errors). Furthermore, the approximation provided by A_n^* is sometimes very close to the values of $\bar{A}_n$. In Cases II and IV, the values of A_n^* are, at the end of the series, significantly smaller than those of $\bar{A}_n$. The approximation is, however, generally a good one even in these cases.

7. TESTING THE SENSITIVITY OF THE BAYES PREDICTION ORDERING POLICY OF THE UPPER ECHELON.

In order to obtain some information concerning the closeness of the Bayes prediction policy to the minimal expected cost policy, we performed the following sensitivity analysis. In each run of the four cases reported in Section 5, we computed the average monthly costs when the Bayesian prediction policy, Z_n^*, was changed to $Z_n^* - 3 + i (i=1, 2, \ldots, 5)$. The averages over the ten replicas of each case were then computed. The results of these computations for Case I and Case II are summarized in Table 6. In this table we present only the averages. The standard errors of these estimates are not significantly different from those given in Table 4. The conclusions that can be drawn from Table 6 are:

(i) At the beginning the average monthly costs associated with $Z_n^* - 2$ are significantly smaller than the other average costs. This trend continues in Case I until the 15th month. In Case II it continues until the 20th months.

(ii) When the ordering policy was $Z_n^* - 1$ we obtain average monthly costs which are significantly larger than those of $Z_n^* - 2$, up to $n=14$ in Case I, and for all $n=1, \ldots, 25$ in Case II.

(iii) The ordering policy Z_n^* is optimal in Case I for all $n=14(1)25$, and in Case II it is not significantly different from that of $Z_n^* - 2$.

TABLE 6. *Average Monthly Total Costs, for Cases I and II, Around the Bayes Prediction Policy*

n	Case I					Case II				
	Z_n^*-2	Z_n^*-1	Z_n^*	Z_n^*+1	Z_n^*+2	Z_n^*-2	Z_n^*-1	Z_n^*	Z_n^*+1	Z_n^*+2
1	283.55	338.50	345.50	352.50	359.50	226.40	280.40	287.40	294.40	301.40
2	199.47	237.40	244.40	251.40	258.40	172.20	208.70	215.70	222.70	229.70
3	172.85	201.13	208.13	215.13	222.13	154.90	182.23	189.23	196.23	203.23
4	142.44	166.86	173.89	180.90	187.90	136.83	165.57	173.07	180.57	187.57
5	126.45	148.98	155.24	162.07	169.07	127.64	157.50	163.70	170.70	177.70
6	117.62	139.83	144.59	150.52	157.28	126.20	155.42	160.75	166.58	172.58
7	111.36	132.90	136.79	141.74	147.26	124.81	151.14	154.71	159.86	165.14
8	108.49	129.44	131.57	136.14	141.92	126.01	150.30	153.93	157.80	162.92
9	104.91	123.89	124.94	129.21	135.13	127.31	149.01	149.34	152.23	157.57
10	106.06	122.80	122.16	125.56	131.21	130.15	148.78	149.68	152.08	156.78
11	106.86	121.75	119.05	121.76	127.46	131.35	148.93	149.47	150.65	155.29
12	108.69	120.65	115.99	119.00	124.84	132.77	149.13	147.55	147.88	153.30
13	109.03	117.06	112.07	115.46	121.39	132.20	147.22	143.53	144.92	150.45
14	110.00	115.60	109.11	112.59	118.59	132.30	146.25	141.75	143.18	148.82
15	109.73	112.95	106.73	110.48	116.55	131.63	143.71	140.11	142.65	148.38
16	109.81	110.76	103.70	107.54	113.67	131.03	141.36	137.04	139.98	145.75
17	111.52	109.99	102.06	105.58	111.76	131.78	140.74	135.44	138.44	144.32
18	115.18	112.19	101.48	104.76	110.98	134.37	142.55	135.77	138.61	144.55
19	118.09	114.43	101.67	104.62	110.89	136.27	144.39	136.08	137.82	143.82
20	119.55	116.06	100.44	103.40	109.70	136.78	144.85	135.55	137.30	143.35
21	121.78	118.41	100.58	102.65	108.99	138.26	146.13	136.04	137.66	143.75
22	123.89	120.68	100.68	102.89	109.26	139.70	147.35	136.67	138.08	144.21
23	125.41	122.27	100.66	102.17	108.57	140.43	147.92	136.57	137.14	143.31
24	127.09	123.91	101.77	102.34	108.76	141.51	148.73	137.19	137.35	143.56
25	129.06	125.89	102.88	102.12	108.56	142.90	150.08	138.00	137.56	143.80

(iv) When Z_n^* is increased to Z_n^*+1 or Z_n^*+2 the average monthly costs increase significantly; however, our computations show that this trend reverses when n is larger than 25.

One can interpret the conclusions drawn from Table 5 not as a necessary indication that the Bayes prediction policy is bad, but as an indication that the prescribed parameter $\gamma=0.85$, according to which Z_n^* was computed, is not a proper one for all n. If γ is decreased Z_n^* is decreased, and when γ is increased so is Z_n^*. We can argue, therefore, that for short periods (n smaller than 15) γ should be chosen smaller than 0.85, and for n greater than 25 γ should be larger than 0.85. The relationship between γ and n seems to be a proper subject for a further study. The application of the Bayes prediction policy with appropriate prescribed parameters, seems to yield desirable results, and, as shown previously, can be simplified by following the previous month demand, that is, by setting $Z_n=X_{n-1}$.

REFERENCES

[1] Fennell, Joseph P., "The Multi-Level Inventory Program—I," The George Washington University Institute for Management Science and Engineering, Technical Memorandum Serial TM−80133 (1971).

[2] Zacks, S., "Bayes Adaptive Control of Two-Echelon Multi-Station Inventory Systems," The George Washington University Institute for Management Science and Engineering, Technical Memorandum Serial TM−61541 (1970).

[3] Zacks, S. and Joseph P. Fennell, "Bayes Adaptive Control of Two-Echelon Multi-Station Inventory Systems, II: Further Development for a Special Case of One-Station Lower Echelon, and Monte Carlo Evaluation, The George Washington University Institute for Management Science and Engineering, Technical Memorandum Serial TM−80141 (1971).

[4] Zacks, S. "On the Optimal Ordering Level of the Lower Echelon in Two-Echelon Inventory Systems," The George Washington University Institute for Management Science and Engineering, Technical Memorandum Serial TM−80237 (1972).

A UNIFIED MODEL FOR DEMAND PREDICTION IN THE CONTEXT OF PROVISIONING AND REPLENISHMENT

Sheldon E. Haber

The George Washington University

and

Rosedith Sitgreaves

Columbia University

ABSTRACT

An improved demand prediction model is presented which incorporates features of two earlier models. The unified model pools usage data classified by repair part class and by component class. The performance of the model is evaluated in a provisioning and replenishment context and compared with that for the current procedure which employs technicians' usage estimates.

0. INTRODUCTION

The problem of demand forecasting for military repair parts is of special importance in the design of efficient inventory systems. This special importance derives from the characteristic that for most items usage is sporadic, i.e., occurs infrequently, if at all, over long periods of time, and low, i.e., demand quantities are generally small when usage is observed.* This suggests the need for designing logistics systems which accommodate to the demand process. This is not to say that other factors, such as those affecting political and military strategy, can be ignored in the structuring of an inventory system. What is implied is that such factors need to be considered within the framework imposed by the demand process.

For example, one consequence of the characteristics of sporadic and low usage is that preferred supply decisions should be made in a probabilistic rather than a deterministic context. With zero or low usage, the application of a deterministic model generally requires that usage estimates be greatly inflated if a sufficient range of repair parts is to be provisioned. Thus, the use of a deterministic model often leads to large over-buys even over the expected design life of the system to be supported. Another illustration in the supply context pertains to the end-consumer echelon. Although the range of items demanded in a given time period is small, it will generally be necessary to stock a large proportion of the installed items if adequate effectiveness is to be achieved. This is so because of the difficulty of predicting which of the large number of items subject to wear will actually fail. Thus, the often followed policy of off-loading stocked items after, say, 1 or 2 years of no observed movement of an item can lead to undesired reductions in readiness. The problem here is not that review and reduction of carried stock at the end-consumer echelon is unwarranted, but that additional information of an item's characteristics, such as its expected usage rate, unit price, unit cube, and military worth, are required before a rational determination can be made of the proper set of items to be off-loaded. Whereas at the end-consumer echelon, sporadic and low usage generally implies that a wide range

*Studies documenting these usage patterns for military repair parts are reviewed in [1].

of parts should be stocked, it also suggests that at resupply echelons a reduced range of parts may suffice. Given the likelihood of long intervals between demands, resupply stock may be augmented and end-consumer shelf stock brought up to desired levels with little risk of decreasing effectiveness (see [4]). From these examples, it is seen that when demand is sporadic and low, the requirement for demand prediction techniques with desirable properties becomes doubly urgent. An approach incorporating the desirable features of two earlier demand prediction models [5] and [6] is the subject of this paper.

As noted, a major focus of the demand prediction problem is the determination of a large but economic range of parts to be stocked at the end-consumer echelon. A second major focus is the determination of the depth, i.e., number of units to be stocked given that a part is carried. Obviously, the quantity of units stocked cannot be large for all items, otherwise investment cost will be large. On the other hand, it is necessary to identify those items for which large quantities are warranted if effectiveness is to be maintained.

Given the intermittent nature of demand, one guideline for improved demand prediction is that past usage should be given as much weight as current usage. Another guideline for improving demand prediction would be to utilize, where possible, information not directly pertaining to the repair item being considered. One application of information not directly related to a given item is provided in [5] where usage data are pooled for repair parts with similar characteristics, in particular, repair parts with the same nomenclature. A second application is described in [6] where information pertaining to the unreliability of the component in which an item is installed is used to modify the usage estimate for the item. It is worth noting that in both of these models, past data are given equal weight with current data.

Experience with the first model indicates that it permits the stocking of a wide range of items without entailing the costs associated with stocking a large number of units for each of the items carried. As noted, however, for some items the stocking of a large number of units is necessary if shortages are to be avoided. This model is somewhat limited in its ability to detect such cases.

The significant feature of the second model is that it permits identification of the small range of parts that are most likely to be demanded in large quantities because they are installed in a small number of components that malfunction relatively often. This is achieved by estimating a numerical measure of the unreliability of the part's parent component. But as structured in [6], the model suffers from the distinct disadvantage that the measure of component unreliability can be applied meaningfully only to components which have malfunctioned one or more times. However, as in the case of repair parts, most components malfunction only rarely. For this model, the problem of intermittent usage still remains, but it manifests itself at the component rather than the repair part level.

Given the two approaches, it appeared desirable to formulate a unified model which incorporated the advantage of each approach taken separately. This new model is described in the next section. It is then evaluated in Sections 2 and 3 using the same context and data employed in the earlier papers.

Besides the specific objective of evaluating the unified model, a second and important objective of this paper is to note some implications of the current, general practice of using technicians' estimates, i.e., estimates provided by supply personnel, for provisioning and replenishment of repair items. The need for technical expertise is particularly important at provisioning when no information specific to a repair item is available. To compensate for this lack of information, the technicians' estimates tend to be conservative because of the need for maintaining adequate levels of effectiveness during the long leadtimes encountered in phasing new material into the supply system. This approach

to provisioning and its extension during replenishment can have an important impact on the structure and manageability of a logistics system. The nature of this impact is discussed in Section 4.

1. A UNIFIED DEMAND PREDICTION MODEL

In describing the unified model, it is useful to outline briefly the two models from which it is constructed. This is done to acquaint the reader with the structure of these models and to facilitate the transition to the unified model.

In the first model [5], classes of repair parts are established on the basis of a specified criterion, for example, nomenclature. For a given part I in the class C, the total quantity of units demanded over T time periods, say, y_I, is assumed to be a random variable with a Poisson distribution whose mean is $\theta_I T$. That is,

$$P(y_I | \theta_I) = \frac{e^{-\theta_I T}(\theta_I T)^{y_I}}{y_I\,!}, \tag{1}$$

where θ_I is the (unknown) parameter of the Poisson distribution for item I in a unit time period.* In this model, θ_I is treated not as a constant, but as the value of a random variable, θ. Values of θ_I for parts in the repair class C are assumed to be generated by a gamma distribution

$$P(\theta|\alpha, \beta) = \left(\frac{\alpha}{\beta}\right) \frac{\theta^{\alpha-1} e^{-(\alpha/\beta)\theta}}{\Gamma(\alpha)}, \tag{2}$$

where the parameters α and β depend on the repair class C. For this (*a priori*) distribution of θ, the Bayes estimator of θ_I with a squared error loss function is

$$E(\theta|y_I) = \frac{T\beta}{\alpha + T\beta} \cdot \frac{\alpha + y_I}{T} \tag{3}$$

for installed parts subject to usage and

$$E(\theta) = \beta \tag{4}$$

for items yet to be introduced into the inventory system and for which usage is not possible.

Estimates of α and β are obtained by the method of moments from the observed distribution of y_I-values for the class C, and the parameter θ_I for part I is estimated as

$$\tilde{\theta}_I = \frac{T\hat{\beta}}{\hat{\alpha} + T\hat{\beta}} \cdot \frac{\hat{\alpha} + y_I}{T} \tag{5}$$

or

$$\tilde{\theta}_I = \hat{\beta}, \tag{6}$$

depending on whether the item is being replenished or provisioned for the first time, respectively.

In the second model [6], two distinct problems are addressed. The notion of component unreliability is explored and a procedure for estimating component unreliability is developed. The measure of component unreliability is then used to estimate an expected demand quantity for a part j in a component v (see (9)). This latter value is the basic building block of the unified model to be described.

To measure component unreliability, it is assumed that the probability of a_{vi} different parts being demanded in component v during patrol i is

*Since the discussion at this point is purposely limited in scope, the reader may wish to review [5], and also [6].

(7) $$P\{a_{vi}/N_v\} = \frac{e^{-\lambda_{vi}}\lambda_{vi}^{a_{vi}}}{a_{vi}!},$$

where N_v is the number of different parts installed in component v and λ_{vi} is the expected number of different parts demanded in component v during patrol i.

Furthermore, it is assumed that λ_{vi} can be decomposed in the following manner:

(8) $$\lambda_{vi} = N_v\theta_{vi} = N_vC_vS_i.$$

where θ_{vi} is the probability that any of the N_v different parts in component v will be demanded during patrol i, S_i is a measure of patrol severity, and C_v denotes the degree to which component v is unreliable. Given this formulation, a procedure for estimating C_v is described in detail in [6].

Having obtained an estimate of C_v, it is used in estimating δ_{vj}, the expected demand quantity for part j in component v. Assuming all $S_i = 1$, this expected value is given by

(9) $$\delta_{vj} = N_{vj}(rC_v)b_j,$$

where C_v is defined as before, r is an appropriately chosen scale factor, N_{vj} is the number of units installed of part j in component v, and b_j is a measure of the replacement rate *per unit of installed population* of part j.

The major innovation of the integrated model is that it applies the methodology of [5] to the problem of estimating C_v. Attention is now directed to this new model which makes possible the estimation of positive values of C_v in (9) for components that have not malfunctioned, or have not been subject to wear because they have not yet been introduced into the logistics system.

In developing the unified model, we begin with the initial assumptions of the second model. That is, we assume that the probability of a_{vi} different parts being demanded in component v during patrol i is given by (7) and that λ_{vi} has the structure given in (8). However, we now consider that component v is one of a family of components, F, established by a given criterion, and that values of C_v for the components belonging to F are values of a random variable whose distribution is a two-parameter gamma distribution of a form similar to (2).

More specifically, we have

(10) $$P(a_{vi}|N_v,C_v,S_i) = \frac{e^{-N_vC_vS_i}(N_vC_vS_i)^{a_{vi}}}{a_{vi}!}.$$

For present purposes, we write

$$y_v = \sum_{i=1}^{T} a_{vi},$$

and, as before, set each $S_i = 1$ so that $\sum_{i=1}^{T} S_i = T$. With the assumption of the independence of the values of a_{vi} in different patrols, we obtain

(11) $$P(y_v|N_v,C_v) = \frac{e^{-N_vC_vT}(N_vC_vT)^{y_v}}{y_v!}.$$

For a given family, F, we consider that the values of C_v for the components in F are generated by a probability distribution,

(12) $$P(C|\alpha,\beta)=\left(\frac{\alpha}{\beta}\right)\frac{C^{\alpha-1}e^{-(\alpha/\beta)C}}{\Gamma(\alpha)},$$

where α and β now represent the parameters of the gamma distribution of C-values for the family F. If we approximate each N_v by a single value N, we can write the joint distribution of the two variables y and C as

(13) $$P(y,\,C|\alpha,\,\beta)=\frac{e^{-NCT}(NCT)^{y}}{y!}\left(\frac{\alpha}{\beta}\right)^{\alpha}\frac{C^{\alpha-1}e^{-(\alpha/\beta)C}}{\Gamma(\alpha)}.$$

The marginal distribution of y is found to be

$$\left[\frac{\alpha}{NT\beta+\alpha}\right]^{\alpha}\frac{\Gamma(y+\alpha)}{\Gamma(\alpha)y!}\left[\frac{NT\beta}{NT\beta+\alpha}\right]^{y}\qquad y=0,\,1,\,2,\,\ldots\,.$$

Estimates of α and β, say $\hat{\alpha}$ and $\hat{\beta}$, are calculated by the method or moments from an observed set of y-values for a family F. These estimates are then used to estimate values of C_v. For replenishment calculations we find

(14) $$C_v=\frac{N_vT\hat{\beta}}{\hat{\alpha}+N_vT\hat{\beta}}\cdot\frac{\hat{\alpha}+y_v}{N_vT}.$$

Note that the approximation of all N_v by a single value N was necessary in obtaining the marginal distribution of y and the estimates $\hat{\alpha}$ and $\hat{\beta}$. In estimating C_v, however, we used N_v, the actual number of different repair parts in component v. For provisioning purposes, we have

(15) $$\tilde{C}_v=\hat{\beta}.$$

From (14) it is seen that in the replenishment calculation, if a positive y-value is recorded for at least one of the components belonging to F, each $\tilde{C}_v$ will be positive even if the associated y_v is zero. Likewise in the provisioning calculation, $\hat{\beta}$ and therefore $\tilde{C}_v$ will be positive.

These estimates are now used in an equation similar to equation (9) to estimate the expected demand quantity for a given part in component v. For the present model, (9) becomes

(16) $$\delta_{vI}=(rC_v)\theta_I,$$

where for the sake of uniformity in notation, j in δ_{vj} is replaced by I and Θ_I replaces N_vb_j.

In applying the unified model in Section 3 to compute stock levels, the following rules were used. First the scale factor r in (16) was chosen in such a manner that the average unreliability coefficient for all components equaled one. In choosing r in this way, the usage rate for repair parts in components of more than average unreliability is increased, while it is decreased for repair parts in components of less than average unreliability. Second, since stock level computations are generally made for a repair part rather than a part-application,* modification of (16) was necessary. The rule adopted was that if a repair part was installed in several different components, the C_v associated with the most unreliable component was used. In view of the difficulty of forecasting demands for repair parts with sporadic and low usage, this rule was chosen in preference to taking a weighted average of C_v-values over all the (component) applications of a part. Adoption of this second rule required that (16) be modified as follows:

*The term part-application refers to a unique part-component combination. If the same part is installed in two different components, this would represent two part-applications.

$$\delta_l = (r \max_r C_r)\theta_l. \tag{17}$$

2. A PRELIMINARY TEST OF THE UNIFIED MODEL

In Table 5 of [6], data were presented which indicated that the probability of an item being demanded in a future time period increased with the unreliability of its parent component as measured during a previous period. The data used to demonstrate this relationship consisted of usage observations for 49,682 part-applications during 82 patrols conducted by Polaris submarines. Data for the first 61 patrols were used to compute a component's unreliability coefficient. The next 21 patrols of data were used to test whether the propensity of a part in a given component to fail was positively related to the component's unreliability coefficient. Of importance for this discussion, components that had not malfunctioned during the base period of 61 patrols—1,160 of 2,391 components fell into this category—had been assigned an unreliability coefficient of zero. With the development of the unified model and the pooling of usage data by component class, it was possible to compute a positive unreliability coefficient for each of these components. Having performed this task, the earlier test was repeated. The results are presented in Table 1.

The format for Table 1 is the same as that for Table 5 in [6]. The top portion of Table 1 pertains to parts for which no usage was recorded for any of their applications during the base period of 61 patrols. The bottom half of the table refers to parts having positive usage in at least one of their applications during the base period. As can be seen from Table 1, parts with zero usage were represented by 38,698 applications, of which 3,005 were in components with unreliability coefficients exceeding 1.0. Similarly, parts with positive usage were represented by only 10,984 applications, of which 3,006 were in components whose unreliability coefficient was greater than 1.0.

The number of demands recorded in the test period of 21 patrols for each of six groups of part-applications is shown in column (2). As can be seen from column (3), there is a marked positive relationship between the proportion of part-applications exhibiting usage in the test period and the unreliability coefficient computed for the parent component using base period data, although the relationship is much stronger for items with usage than for items with no usage in the base period.

From the figures in Table 1, the important finding to be noted is that the adoption of the approach developed in [5] to the estimation of component unreliability factors preserves the relationship between component unreliability and repair part usage found in [6].

3. DEMAND PREDICTION IN THE CONTEXT OF PROVISIONING AND REPLENISHMENT

Since a primary context of the demand prediction problem is that of provisioning and replenishment, it is necessary and useful to evaluate the unified model in the environment in which it is to be applied. Since an actual, controlled experiment could not be performed, historical data were employed in a simulation exercise.

Usage data were available for 21,225 repair parts for a base period of 61 patrols and a later period of 21 patrols. Since all estimates of demand prediction parameters, i.e., the parameters $\hat{\alpha}$ and $\hat{\beta}$ in (5), (6), (14), and (15), were computed using the data for the first 61 patrols, the subsequent 21 patrols were designated as the period for which replenishment was to be accomplished. Also available were additional data covering 35 patrols for a smaller set of 4,094 repair parts. Each of these items was a "new" item in that it was not included in the larger set of repair parts. Moreover, no component in the smaller set was also contained in the larger set, i.e., the components in the smaller and larger sets of

TABLE 1

Component unreliability coefficient[a]	Number of part-applications[b] (1)	Number of part-applications demanded[c] (2)	Percent of part-applications demanded $\frac{(2)}{(1)} \times 100$ (3)
Parts with zero usage:			
$0 < C_v^* \leq 0.1$	20,229	61	0.3
$0.1 < C_v^* \leq 1.0$	15,464	241	1.6
$1.0 < C_v^*$	3,005	125	4.2
Total..	38,698	427	1.1
Parts with positive usage:			
$0 \leq C_v^* \leq 0.1$	3,450	30	0.9
$0.1 < C_v^* \leq 1.0$	4,528	501	11.1
$1.0 < C_v^*$	3,006	1,525	50.7
Total..	10,984	2,056	18.7

[a] The scaled coefficient rC_v is here denoted by C_v^*.
[b] Based on patrols 1–61.
[c] Based on patrols 62–82.

data were mutually exclusive. For these reasons, we considered the period encompassing the 35 patrols as the period for which provisioning was to be accomplished. For provisioning, the data base used in estimating $\hat{\alpha}$ and $\hat{\beta}$ values was the same as the one employed for replenishment, namely, the first 61 patrols of data for the larger set of items.

Although the provisioning period of 35 patrols and replenishment period of 21 patrols refer to different groups of items, it is useful to consider the provisioning and replenishment calculations as being performed for the same set of items. As indicated below, in gross terms the two sets of items appear to be similar. The advantage of doing this is that it makes possible a rough assessment of the demand prediction problem through a provisioning-replenishment cycle. Additionally, given the difficulty of the demand prediction problem, it permits some insight into possible alternative strategies for designing an inventory system.

The approach employed in accomplishing these objectives was to develop stockage lists which differed in only one respect. The stock quantities for each list were computed using a single inventory model, that described in [3], and all input data were exactly the same* except the repair item usage rates which were estimated using the demand prediction procedures described below. Although we are concerned with the effect of different demand prediction techniques on the structure and manageability of an inventory system, the stockage lists are restricted to the end-consumer, in this case, Polaris submarines. However, the findings would be applicable to other echelons as well.

* Inventory model parameters used in computing all stock quantities are found in [8].

Four stockage lists were compared by matching item stock quantities against item demand quantities during the provisioning and replenishment periods and then calculating shortage counts. The first stockage list, called the Technicians' Estimate Model and denoted as Model I, uses technicians' usage estimates which were obtained from Navy files. The second, Model IA, is a modified version of Model I; discussion of Model IA is temporarily deferred. The third, called the Pooled Repair Part Model, is described in [5]: the particular version applied here is Model IIA (see [5]) in which usage data is pooled by individual repair part class. As previously indicated, unlike the unified model, in the Pooled Repair Part Model no attempt is made to utilize component data to modify individual repair part usage rates. The fourth, Model III, is the unified model developed in this paper; the repair part usage estimates, $\tilde{\theta}_l$, are the same as in the Pooled Repair Part Model IIA, but in the unified model they are multiplied by $r \max_r \tilde{C}_r$.†

The component unreliability coefficients, $\tilde{C}_r$, were estimated only one time, at the end of patrol 61. Once computed they were treated as fixed parameters in all provisioning and replenishment stock list computations. Item usage estimates, θ_l, on the other hand, were updated for Models IIA and III after each patrol during provisioning and replenishment, based on the usage y_l accumulated to that patrol. Updating of the models based on the technicians' usage estimates was not performed since this technique provides no guidelines as to how this is to be done.

TABLE 2. *Stockage List Characteristics*

Provisioning and replenishment	Models			
	Tech. est. I	Modified tech. est. IA	Pooled repair parts IIA	Unified III
Provisioning[a]				
Range	3.9	3.8	3.7	3.7
Depth	18.9	7.7	4.4	8.9
Storage space	489.3	214.1	189.5	202.5
Dollar value	450.2	229.9	231.8	276.0
Replenishment[b]				
Range	18.6	15.2	18.9	13.3
Depth	112.3	39.5	25.4	28.2
Storage space	1,953.5	778.0	534.2	596.8
Dollar value	2,703.1	1,000.2	960.4	924.7

[a] Averages for patrols 1–35, for 4,094 items. All figures in thousands.
[b] Averages for patrols 62–82, for 21,225 items. All figures in thousands.

The stockage lists corresponding to each demand prediction technique are shown in Table 2.* Of immediate interest is that the range (number of different items stocked), depth (number of units

† To compute the unreliability coefficients, $\tilde{C}_r$, components were grouped into 149 classes. A difficulty in classifying component classes needs to be mentioned. Components managed by the Strategic Systems Projects Office (SSPO) had to be classified by major system, e.g., navigation. All other components were classified in terms of nomenclature, e.g., pumps, compressors, etc. Because of this, there were cases where a given design entity was classified using both classification criteria. Additionally, with the SSPO systems there were a large number of components that were very dissimilar in design and function. These difficulties in classification reduce the predictive capability of the unified model.

*With the exception of the figures pertaining to storage space, which is measured in cubic feet, all figures in Table 2 for Models I and IIA are taken from [5]. The figures for Models I and IIA in Tables 4 and 5, and the top half of Table 6, are also taken from this source.

stocked), storage space, and dollar value of stock for Model I for the replenishment period are approximately five times the corresponding values for the provisioning period. Thus, in aggregate terms, the composition of the 21,225 replenishment items appears to be similar to that for the 4,094 provisioned items. Similarities between the two periods is also seen from the demand data in Table 3. Both the range and depth of demand during the replenishment period are again approximately five times as large as the corresponding figures for the provisioning period.

From a substantive point of view, important differences between Model I, incorporating the technicians' usage estimates, and the unified model, Model III, are apparent. These are summarized by dividing the figures in the last column of Table 2 by those in the first column:

	Model III/Model I	
	Provisioning	Replenishment
Range	0.95	0.72
Depth	.47	.25
Storage space	.41	.31
Dollar value	.61	.31

One sees from these ratios that Model I costs more than Model III and that the differential in investment cost is much larger for replenishment than for provisioning. It should be noted that the expensiveness of Model I is not limited to dollar value, but extends to the range of items stocked, particularly for replenishment; the quantity of units stocked; and the space required for storing stocked units. These latter types of costs are particularly important for flexibility in echeloning repair items, e.g., in determining the feasibility of advanced land-based storage sites; in improving the distribution of space among alternative uses, including the possibility of a net reduction in space requirements; in reducing the need for management resources; and in permitting improved management, such as record keeping and control of a smaller volume of material. That the differentials in favor of Model III are larger for replenishment than provisioning is also of some interest. The larger differentials are explainable in terms of the lesser availability of information at provisioning which increases the number of units needed to achieve a given assurance against stockout for Model III. This is not the case for Model I where the degree of expertise presumed for the technician is assumed independent of current demand experience.

TABLE 3. *Range and Depth of Demanded Items*

	Number of different items installed	Average number of items demanded	
		Range[a]	Depth[b]
	(1)	(2)	(3)
Provisioning	4,094	23.5	86.3
Replenishment	21,225	95.3	485.6

[a]Number of different repair parts demanded.
[b]Number of units demanded over all demanded parts.

A second point noted earlier is that relative to the range and depth of demanded items, the range and depth of stocked items is very large. One implication of this is that given sporadic and low demand,

improvements are more likely to be achieved by reducing the cost of attaining a given level of effectiveness than in reducing shortages of material. Evidence supporting this conclusion is found in Tables 4 and 5.

TABLE 4. *Range Shortage Counts*

Provisioning and replenishment	Models			
	Tech. est. I	Modified tech. est. IA	Pooled repair parts IIA	Unified III
Provisioning[a]				
Items not stocked	1.0	1.8	0.9	1.0
Items stocked	3.6	6.1	6.7	6.2
All items	4.6	7.9	7.6	7.2
Replenishment[b]				
Items not stocked	2.5	8.3	3.0	4.5
Items stocked	19.5	32.1	23.1	14.6
All items	22.0	40.4	26.1	19.1

[a] Averages for patrols 1–35, for 4,094 items.
[b] Averages for patrols 62–82, for 21,225 items.

TABLE 5. *Depth Shortage Counts*

Provisioning and replenishment	Models			
	Tech. est. I	Modified tech. est. IA	Pooled repair parts IIA	Unified III
Provisioning[a]				
Items not stocked	1.7	2.7	1.5	4.4
Items stocked	33.8	47.7	48.4	44.4
All items	35.5	50.4	49.9	48.8
Replenishment[b]				
Items not stocked	6.8	27.9	4.8	7.6
Items stocked	172.3	262.3	225.0	143.3
All items	179.1	290.2	229.8	150.9

[a] Averages for patrols 1–35, for 4,094 items.
[b] Averages for patrols 62–82, for 21,225 items.

Tables 4 and 5 show the average range and depth of shortages per patrol during the provisioning and replenishment periods. The range measure indicates the number of different items for which the quantity of units demanded during a patrol is more than the quantity stocked. The depth measure indicates the difference between the number of units demanded and the number of units stocked, summed over all items for which a shortage was experienced during a patrol. In these tables, the shortage counts are shown separately for stocked and nonstocked items. By distinguishing these categories, some measure of the difficulty of the demand prediction problem can be obtained. For example, between 13 and 28 percent of all range shortages during the provisioning period originated among items not carried, even though the percentage of installed items in this category was no more than 10 percent for any model. Of particular interest, relatively few shortages were registered among

not carried items by Model III during the replenishment period, even though this model stocked as little as 63 percent of the repair parts.

In comparing Models I, IIA, and III in terms of range and depth shortages, it is seen that Model I performed best during provisioning. During replenishment it also performed well vis-a-vis the other models, but it did not perform as well as Model III. Comparing Models IIA and III, one notes that these two performed equally well during provisioning. During replenishment, however, range and depth shortages were lower for the latter. The same results were obtained, as can be seen from Table 6, when the repair items in the sample were classified by military essentiality.*

TABLE 6. *Range and Depth Shortages for Highest and High Essentiality Items*

Provisioning and replenishment[a]	Models			
	Tech. est. I	Modified tech. est. IA	Pooled repair parts IIA	Unified III
Highest essentiality items:				
Provisioning[b]				
Range shortages	0.2	0.2	0.4	0.5
Depth shortages	1.9	3.4	5.1	5.3
Replenishment[c]				
Range shortages	2.2	3.7	2.4	1.4
Depth shortages	19.1	29.7	23.3	19.7
High essentiality items:				
Provisioning[b]				
Range shortages	0	0	0	0
Depth shortages	0	0	0	0
Replenishment[c]				
Range shortages	6.8	13.9	8.1	5.1
Depth shortages	92.9	161.0	127.5	70.8

[a]Includes stocked items and non-stocked items.
[b]Averages for patrols 1–35, for 4,094 items.
[c]Averages for patrols 62–82, for 21,225 items.

As just mentioned, range and depth shortage counts were similar for Models I and III: they were smaller during provisioning for the former, but during replenishment they were smaller for the latter. In terms of costs, however, they were very dissimilar. An important question, therefore, is whether Model I could be improved by scaling the technicians' estimates so as to reduce investment costs. One difficulty with this is that in practice there is no benchmark to which to scale the technicians' estimates. This was overcome here by reducing the technicians' estimates until the dollar value of Model I was in the neighborhood of that for Models IIA and III. The across-the-board reduction in technicians' estimates necessary to accomplish this was one-sixth. This approach is denoted as Model IA.

In terms of range and depth of items stocked, Model IA most closely resembles Model III. As can be seen from Tables 4–6, for the provisioning period, the shortage counts for Model IA were higher than for Model I, but about equal to Model III. On the basis of these results it would appear that in the absence of knowledge of an item's failure rate, as prevails during provisioning, technicians do exhibit expertise. However, although the modified technicians' estimates compared favorably with those com-

*The specific methodology used for determining an item's military essentiality is described in [2].

puted using Model III during provisioning, this was not the case for replenishment. For replenishment, the differences in range and depth shortages were substantial and Model III was clearly superior.

With respect to Models IIA and III, the shortage counts were similar during the provisioning period, but they were higher for the former model than for the latter during replenishment. Also of importance, the reduction in shortages during the latter period was realized despite the much smaller range of items carried by Model III. The data thus substantiate the advantage cited for the unified model, namely, that it identifies items that should be stocked in above average quantities. The successful accomplishment of this objective is evidenced by a marked redistribution of stock quantities among a small range of items, while at the same time effecting a reduction in shortages.

In reviewing the empirical findings in Tables 4–6, a question which comes to mind is why Model III perfomed better than Model IIA during replenishment, but only as well as the latter during provisioning? The most likely explanation is that, as noted earlier, the components for which repair parts were provisioned were mutually exclusive of the components for which repair parts were replenished. On the other hand, the latter components were the same as those for which usage data were collected during the 61 patrol base period. In this aspect, the simulation may have provided a stricter test of Model III during provisioning than real life, since new systems being provisioned very often are comprised of many components that had been installed in earlier configurations of the system. Had there been some commonality of base-period components and provisioning period components, the shortage counts for Model III would most probably have been lower.

It also is likely that the predictive capability of Model III was constrained in other ways which may have led to an upward bias in its shortage counts for the replenishment as well as the provisioning period. First, the base period of 61 patrols was extremely small, representing less than one-half year of usage for all Polaris submarines. Second, for Polaris submarines the percentage of repair items with positive usage appears to be smaller than for other types of ships and aircraft* (see [1]). Finally, the classification of components by class presented difficulties which would be absent for most other types of ships and aircraft.

4. IMPLICATIONS OF THE STUDY FOR INVENTORY MANAGEMENT

In this section, some additional observations are presented based on the analysis of the previous section. These observations are of a general nature in that they do not pertain to any one demand prediction technique. Rather, they relate to general policy considerations in inventory management.

One consideration that is highlighted by the figures in Table 2 is the need for compatibility of demand prediction techniques between the provisioning period and subsequent replenishment periods. For example, the strategy of relying on technicians' usage estimates, i.e., Model I, for provisioning and the unified model for replenishment, because each of these techniques performs best during the period in which it would be applied, can lead to incompatible inventory systems. Under this strategy the capacity of the inventory system needed to satisfy provisioning requirements would be larger than the capacity needed to satisfy replenishment requirements.

Another consideration pertains to changes in the structure of an inventory system over the life of the system being supported, given that the same demand prediction technique is used for both provisioning and replenishment. Again referring to Table 2, if Model III were used for provisioning and replenishment, as usage information is accumulated the range of items managed and the space needed

---

*These limitations may have also reduced the preductive capability of Model IIA.

for storage would likely decline. This has important implications for requirements and management of resources other than repair items. It also raises the question of whether inventory systems should be designed as static or dynamic structures, and if the former whether the appropriate structure is the one best suited to provisioning or to replenishment.

Assuming a static structure, and looking at the full life of a given system, it would appear that inventory systems ought to be designed for the replenishment period. This is the strategy implicit to deferred procurement. A penalty of its adoption, however, is the likelihood of increased shortages during provisioning. To some extent the risk of reduced effectiveness can be lessened via changes in the mix of items stocked over the provisioning-replenishment cycle. For example, during provisioning greater weight might be given to items with higher essentiality. Over time, failure to supply low essentiality items can lead to accumulated reductions in effectiveness, and the weight given to this class of items would need to be increased. The problem of determining the proper mix of repair items over time would be minimized by the availability of accumulated usage data and its utilization in an effective manner.

Finally, the difficulty of the demand prediction problem should be emphasized once again. This is clearly evident in comparing the figures in Tables 4 and 5 with those in Table 3. As can be seen from these tables, the supply quantity was less than the demand quantity for 20 percent of the repair items during the replenishment period (using Model III) and the provisioning period (using Model I).* Although it is felt that the unified model is a distinct advance in the state of the art of demand prediction for military repair parts, its primary contribution appears to be in reducing the costs of attaining a given level of supply effectiveness. Marked improvement in supply effectiveness, itself, could not be demonstrated, perhaps because of the data and other limitations noted above, so this objective still remains to be achieved.

5. SUMMARY

In [5] and [6] two distinct aspects of the demand prediction problem for military repair parts were investigated. In both cases accumulated information not directly pertaining to the item being considered was used in the estimation of the part's usage rate. The first [5] dealt with the difficulty of estimating repair part usage rates when usage is sporadic and low; in particular, when no usage is experienced for an item over long periods of time. The methodology for accomplishing this was to pool usage data by repair part class. The second [6] pertained to the problem of identifying unreliable components and using this information to modify repair part usage estimates. The first model focused on the range of items to be stocked while the second one focused on the depth of units to be stocked. In this paper a unified model is presented which incorporates the advantages of both of the earlier ones. This is done by applying the methodology in [5] to the problem discussed in [6], i.e., the estimation of component unreliability factors.

The unified model is compared with the current procedure of utilizing technicians' usage estimates for computation of stockage lists. The context of the study is that of provisioning and replenishment. Although the unified model by no means solves the demand prediction problem, it is shown, despite a number of simplifying assumptions, to provide a means for substantially reducing the costs of inventory

*This suggests the need for inventory rules which take precedence over the quantity stocked on the basis of usage data. Examples of such rules, e.g., the stocking of at least one unit of every high and highest essentiality item regardless of its usage rate, are found in [7] and [8].

management. The reduction in costs is not only in terms of investment of stock, but in other important nonmaterial costs as well.

Examination of the demand prediction problem in the context of provisioning and replenishment also provides some insight into alternative strategies for structuring inventory systems. In particular, the structure of the inventory system most appropriate for provisioning appears to be different from the structure appropriate for replenishment. Changes in structure may be warranted for reasons other than those pertaining to demand prediction, such as the phasing out of a system being supported. However, changes in structure may also be warranted if only because the efficient use of accumulated usage data permits improved estimation of requirements over time. As might be expected, it was found that as usage data was accumulated, the range and depth of carried stock diminished appreciably. For a static inventory system, this raises the important question of whether it should be structured to satisfy the requirements associated with provisioning or with replenishment. This and associated problems require further investigation.

6. ACKNOWLEDGMENT

The authors wish to thank William Hise and Joseph D'Amalio for their programming assistance.

REFERENCES

[1] Astrachan, Max and Albert S. Cahn (editors), "Proceedings of Rand's Demand Prediction Conference, January 25–26, 1962," The Rand Corporation, Santa Monica, California, RM–3358–PR (1963).

[2] Denicoff, M., J. Fennell, S. E. Haber, W. H. Marlow, F. W. Segel, and Henry Solomon, "The Polaris Military Essentiality System," Nav. Res. Log. Quart. **11**, 235–257 (1964).

[3] Denicoff, M., J. Fennell, S. E. Haber, W. H. Marlow, and Henry Solomon, "A Polaris Logistics Model," Nav. Res. Log. Quart. **11**, 259–272 (1964).

[4] Haber, Sheldon E., "Simulation of Multi-Echelon Macro-Inventory Policies," Nav. Res. Log. Quart. **18**, 119–134 (1971).

[5] Haber, Sheldon E. and Rosedith Sitgreaves, "A Methodology for Estimating Expected Usage of Repair Parts with Application to Parts with No Usage History," Nav. Res. Log. Quart. **17**, 535–546 (1970).

[6] Haber, Sheldon E., Rosedith Sitgreaves, and Henry Solomon, "A Demand Prediction Technique for Items in Military Inventory Systems," Nav. Res. Log. Quart. **16**, 302–308 (1969).

[7] Solomon, Henry and Marvin Denicoff, "Simulations of Alternative Allowance List Policies," Nav. Res. Log. Quart. **7**, 137–149 (1960).

[8] U.S. Navy Special Projects Office, "Repair Parts Support for Special Projects Office Fleet Ballistic Missile Equipments," Instruction P4423.27A (22 May 1967), Washington, D.C.

IMPACT OF AN ALL VOLUNTEER FORCE UPON THE NAVY IN THE 1972–1973 TIMEFRAME

A. S. Rhode

Office of the Chief of Naval Operations
Department of the Navy

J. J. Gelke, Cdr., U.S.N.

and

F. X. Cook

Mathematica, Inc.

ABSTRACT

This paper develops estimates of true volunteer levels for 1972 and 1973, based on experience gained through 1970 draft lottery data. The paper also formulates estimates of the qualitative characteristics of a 1972–1973 Navy volunteer force, and establishes a relationship between rate of volunteerism and military pay. Utilizing estimates generated in the paper, Navy military personnel budget requirements for FY '72 and '73 are presented.

INTRODUCTION

The major objectives of this study are to develop estimates of MPN (Military Personnel, Navy) budget requirements for FY '72* and FY '73, under a draft-free environment, and to describe the "quality" characteristics of a future Navy volunteer force.

These objectives are addressed in three parts. Part I develops levels of true volunteers for 1970, through analysis of draft lottery data. Projections are then made to present estimates of the levels of volunteers, which can be expected in 1972 and 1973, in the absence of a draft. Part II, through regression analysis, establishes a relationship between rate of true volunteerism, and military and civilian pay. Part III explores the implication of varying military pay considerations upon the number and educational characteristics of a volunteer force.

The regression analysis techniques utilized in this paper are similar to those used in previous study efforts† in this area. However, these previous studies were undertaken prior to the institution of the draft lottery system and thus relied heavily upon attitude surveys to determine levels of true volunteerism. Additionally, previous efforts had no objective method of determining characteristics of true volunteers, such as their educational attainment or general mental makeup. This paper, as previously reported, utilizes an objective approach whereby draft lottery data are analyzed to determine measures of true volunteerism, and characteristics of the true volunteer group.

The study developed that approximately 55 percent of the Navy enlistees, who have been assigned draft lottery numbers (i.e., those born between 1944 and 1951), were true volunteers. Shown below is a

---

*SECDEF has stated that the United States will go to an all volunteer force concept by mid 1973; however, since present draft laws are due to expire in 1972, the possibility exists that the volunteer concept may have to be instituted in FY '72.

†See Ref. [1–3, 5, 6].

compilation of "optimistic" and "pessimistic" estimates of "true" volunteers projected for 1972 and 1973 as developed in this paper.

Projected Estimates of True Volunteers for All Services, 1972–1973

Estimates	Army		Navy		Marines		Air Force	
	'72	'73	'72	'73	'72	'73	'72	'73
Optimistic	104,000	106,000	69,000	70,000	53,000	54,000	47,000	48,000
Pessimistic	82,000	84,000	54,000	56,000	28,000	28,000	32,000	33,000

The elasticity of supply of USN true volunteers with respect to annual pay is estimated to be 0.9–1.5, with a most likely value of 1.2.

Forecasted USN accessions requirements are 85,000 and 135,000 for FY '72 and '73, respectively. Forecasted annual numbers of true volunteers for these same years are 55,000–70,000, in the absence of any pay increase.

It appears that 85,000 annual USN accessions are possible in a draft-free environment in FY '72 and '73, assuming that the Navy is willing to enlist a greater number of lower mental category volunteers. The FY '70 percentage of accessions in mental categories I, II, and IIIa was 70 percent. If the Navy were to fulfill 85,000 true-volunteer accessions in FY '72 or '73 with no pay increase, this percentage would decrease to 40–55 percent (similar to 1955–56 levels). In addition, this paper indicates that it is highly unlikely that 135,000 true-volunteer accessions could be attained without a pay increase, regardless of the mental profile of the enlistments.

Estimated first-term base pay increases and resultant incremental MPN expenditures are shown below. These estimates are based on the assumption that short-term Navy accession requirements determine necessary pay increases.

USN accessions required	Increase in first-term base pay required (%)	Estimated total incremental MPN expenditures FY '72 or '73 (millions constant 1970 dollars)
85,000	23–51	200–430
135,000	83–128	690–1060

The above expenditures would be necessary to attain required numbers of accessions and maintain 70 percent of total enlistments in mental categories I, II, and IIIa. Military Personnel, Navy estimates are total incremental amounts for all USN enlisted personnel based on the following assumptions:

(1) FY '72 Enlisted Force Strength Plan ("B" Mod) applies.

(2) FY '72 or '73 incremental MPN expenditures for enlisted grades E–1 through E–3 will be $30 million per 10 percent increase in first-term base pay (per BUPERS cost model).

(3) FY '72 or '73 incremental MPN expenditures for enlisted grades E–4 through E–9 are based upon pay increases which maintain a suitable "payline" across grades (see Ref. [7, Annex E] for details).

The Army faces potentially the greatest difficulty in attaining required accession levels in a draft-free environment. Thus, Army requirements may be the driving force in determining DOD-wide military

pay scales for an all-volunteer force. Assuming this to be the case, the required pay increases and resultant impact on Navy MPN expenditures would be as follows:

Required Army accessions	Increase in first-term base pay required (%)	Estimated total incremental MPN expenditures (millions constant 1970 dollars)
180,000	45–80	380–670
200,000	59–98	500–810
220,000	73–117	610–970

Pay increases required by the Army exceed those required to attract 85,000 USN true volunteers in FY '72, but fall below USN requirements for FY '73. The net effect of basing pay increases on Army needs would be to shift the mental distribution of FY '72 USN enlistees (85,000) upward (85–100 percent of total enlistments in mental categories I, II, and IIIa) and to shift the mental distribution of FY '73 USN enlistees (135,000) downward (53–65 percent of total enlistments in mental categories I, II, and IIIa).

DETERMINATION OF TRUE VOLUNTEERS FOR THE NAVY AND OTHER SERVICES FOR 1970 AND OUT YEARS 1972–73

INTRODUCTION.

Reinstatement of the draft lottery system in January 1970 has facilitated the use of a quantitative approach in the determination of the draft motivation of volunteers. It is felt that young men who enter the lottery develop perception of their draft vulnerability based on their relative standing in the manpower pool and that decisions to enlist are affected by this perception. Levels of "true" volunteers are developed for all services for 1970 based upon current lottery data, and projections are made estimating the volunteer levels in 1972 and 1973.

The general approach to determining levels of volunteers for 1970 is broken into three steps. Step one develops a level of true volunteers from those enlistees who are presently in the draft pool (i.e., all young men born between 1 January 1944 and 31 December 1950). Step two develops levels of true volunteers for those enlistees who do not enter the draft pool until 1 January 1971, but who were assigned lottery numbers in the 1 July 1970 drawing (i.e., those with birthdates between 1 January and 31 December 1951). Step three deals with those young men who are not yet affected by the draft lottery, that is, these who have not yet been assigned lottery numbers (i.e., those born between 1 January 1952 and 31 December 1953). These three estimates when totaled, represent levels of true volunteers for calendar year 1970. Based upon these 1970 estimates, volunteer projections are made for 1972 and 1973, for all services.

THE DRAFT LOTTERY

On 1 December 1969, the draft lottery was reinstated, after a lapse of 28 years. The lottery affected all young men who had birthdates between 1 January 1944 and 31 December 1950, and who had not previously served in the military. The chance drawing, in essence, determined which young men of approximately 850,000 would be drafted into military service and which would be left free from call. Draft lottery numbers from 1 to 366 were assigned according to when a man's birthdate was drawn

from a bowl containing all possible birthdates. The first date drawn was 14 September, meaning that every man between 19 and 26 whose birthday fell on that date was given the draft lottery number 1. Those with an April 24th date were assigned number 2, and so on through number 366 which was assigned to the birthdate 8 June. At this time, government officials stated that as a general rule men drawing the lowest third of the lottery numbers, 1 through 122, could be certain that they would be drafted (starting 1 January 1970), that men with the highest third of the numbers, 244 through 366, could be assured that they would not be drafted and that those in the middle, 123 through 243, would be uncertain throughout the year. This information was widely publicized, and it is felt that as a result, young men apparently formed perceptions of their vulnerability to the draft [8, 9]. This expectation will hereafter be referred to as Perceived Draft Vulnerability. On 1 July 1970, the second lottery was held to assign numbers to those with birthdates in 1951. Although this birth group was assigned lottery numbers on 1 July, they did not actually enter the pool until 1 January 1971 [10].

APPROACH FOR DETERMINING TRUE VOLUNTEER RATE

As previously stated, young men in the draft lottery have developed their own *Perceived Draft Vulnerability*. That is to say they have apparently been able to judge their chances to be drafted as: certain to be drafted, uncertain, or certain not to be drafted; depending on their lottery number. Statistical data obtained from the Army Recruiting Command presents numbers of volunteers entering the Navy and other Services by *Lottery Category*.* Lottery Categories are numbered I through XVIII and each contains 20 lottery numbers, except Category XVIII which contains 26. Category I is composed of those most vulnerable to the draft, those with lottery numbers 1 through 20. Category II contains lottery numbers 21 through 40, and so on through Category XVIII which contains lottery numbers 341 through 366. Annex A of Ref. [7] is a compilation of the recruiting statistics for the period 1 January 1970 through 31 October 1970. Figure 1, derived from these data, presents a measure of recruits entering each of the services by lottery group with areas A, B, and C representing the three general categories of Perceived Draft Vulnerability:

A—Those certain to be drafted (Lottery Numbers 1–122)

B—Uncertain (Lottery Numbers 123–243)

C—Those certain not to be drafted (Lottery Numbers 244–366)

Moving to the right, enlistments tend to decrease as draft vulnerability decreases, and a leveling effect occurs almost upon reaching the "C" boundary (certain not to be drafted). This leveling is significant and represents what is considered to be the apparent level of *true volunteers* for those draft eligible men who entered the Navy between 1 January and 31 October 1970. In the following section the estimated levels of true volunteers for the entire year 1970 and for 1972 and 1973 are projected.

VOLUNTEER PROJECTIONS

As mentioned above, the average number of enlistees in the lottery groups bounded by the "C" draft lottery category (certain not to be drafted) has been accepted as representative of the level of true volunteers. Thus the total number of "true" volunteers for the period 1 January through 31 October 1970, are derived to be:

---

*U.S. Army Recruiting Command, *HQUSAREC Form 1136, Monthly Report Supplement to the Qualitative Distribution Report by State and Random Sequence Birthdate Groupings* (through October 1970).

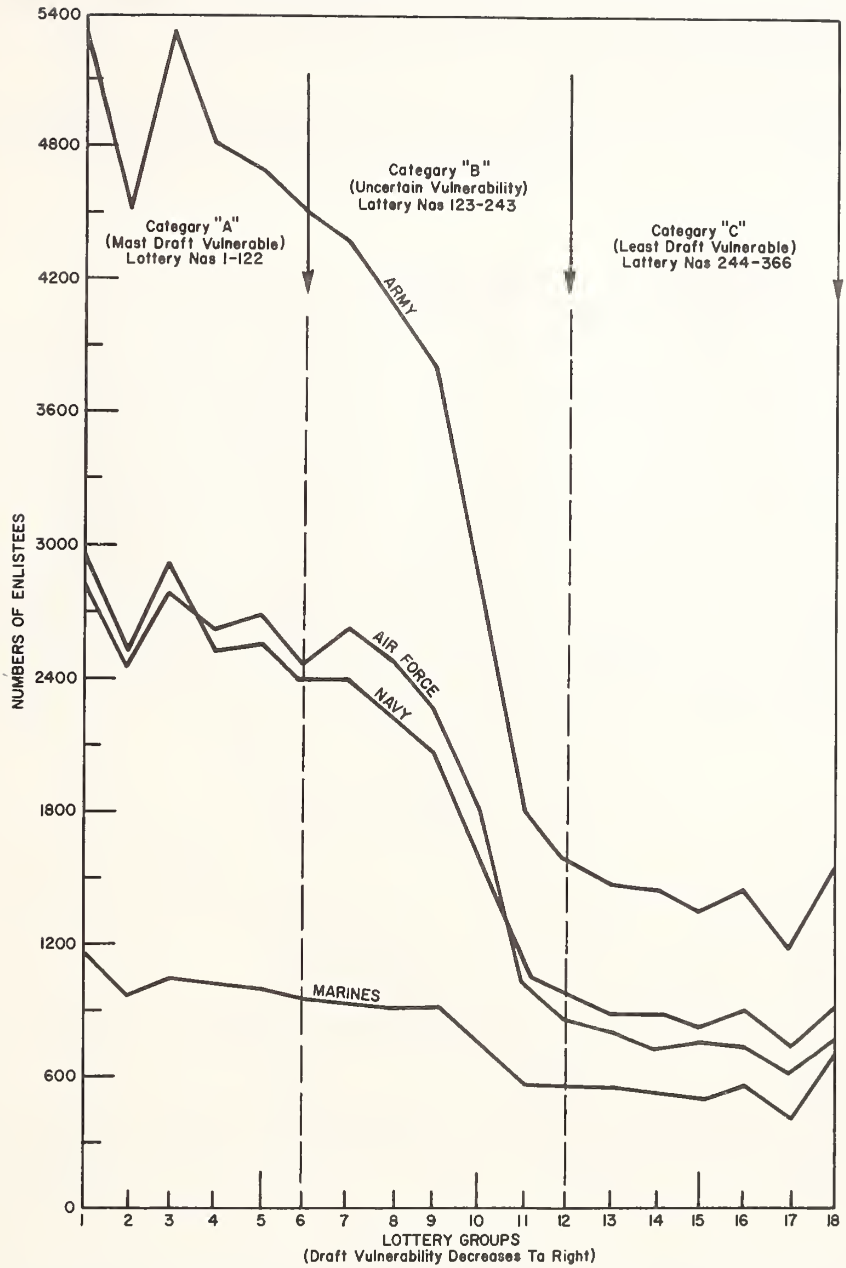


FIGURE 1. Draft lottery groups (draft vulnerability) vs numbers of enlistees/group (birth group 1944–1950).

$$V_1 = \left(1/n \sum_{i=13}^{17} v_i\right) N^* = 15{,}170,$$

Where V_1 = Total true volunteers for the period 1 January through 31 October 1970 with birthdates *1944–50,*

n = Number of lottery groups,

---

*Groups XIII through XVII are chosen since they represent the "true" volunteer level (see Fig. 1). Group XVIII is not average since it contains a distortion caused by the fact that it contains 26 lottery numbers as compared with 20 for all other groups.

v_i = Number of volunteers per lottery group (birth groups 1944–50) (see Ref. [7, Annex A] Navy Enlistments), and

N = Total number of lottery groups.

In projecting these estimates, seasonal variations are ignored on the assumption that the remaining months' volunteers mean will not vary excessively from the 10-month mean.

Thus:

$$V_f = 18{,}200,$$

Where, V_f = Total true volunteers for 1970 with birthdates 1944–50.

During the same January to October period, the Navy also enlisted several thousand first term men who had not yet entered the draft pool* (those with birthdates between 1951 and 1953). Between January and June 1970, young men with the birth year of 1951 had no definite means on which to estimate their future draft vulnerability. However, on 1 July 1970, this group was assigned lottery numbers, and it was indicated at the time that they would enter the draft pool on 1 January 1971. As a result of the July drawing, these young men could begin to perceive their draft vulnerability in 1971 [10].

Data contained in Annex A of Ref. [7] and displayed in Figure 2 indicate that a portion of the 1951 birth year group has apparently already begun to enlist on future expectations of being drafted. Thus after 1 July 1970, the 1951 group apparently began reacting to draft pressure in a manner similar to that of the 1944–50 group. (See Fig. 1.) This additional level of true volunteers for the July through December time frame was estimated using previous methodology:

$$V_2 = \left(1/n \sum_{i=13}^{17} v_i\right) N = 6{,}066,$$

Where V_2 = Total true volunteers for 1951 birth group for period 1 July thru 30 October 1970.

*Draft pool is defined as being the reservoir from which inductees are called.

Again ignoring seasonal variation, the total true volunteers (birth year 1951) for the period 1 July through 31 December 1970 was computed to be 9100.

The remaining true volunteers, those who are not as yet in the lottery or who have not been assigned lottery numbers, are those young men who were born in 1952 and 1953, and those born in 1951 who entered prior to the second lottery drawing on 1 July 1970. The conclusion might be drawn that these men were all true volunteers since they enlisted under no apparent draft pressure. It is felt, however, that to say this is overstating the case. When dealing with young men who have not yet been assigned draft lottery numbers, a new factor emerges which will be termed Perceived *Future Draft Vulnerability*. That is to say there are apparently a certain number of enlistees (not yet in lottery) who volunteer on the assumption that they will probably receive a sufficiently low lottery number to be drafted, when they actually enter the lottery. It is difficult to place a numerical value on the percentage of enlistees falling into this category. Some planners utilize the results of a 1968 attitude survey conducted by DOD which found that about 35 percent of those questioned (between 17 and 18) were draft motivated.* It is felt, however, that this estimate is overstated, in that today's draft pressures are not the same as they

*Office of SECDEF (M&RA) apply average factors based upon an attitude study conducted by DOD in 1968. This survey found that among 17 and 18 year olds about 37 percent of Army enlistees were draft motivated, 36 percent of Navy enlistees, 24 percent of Marines, and 33 percent of Air Force.

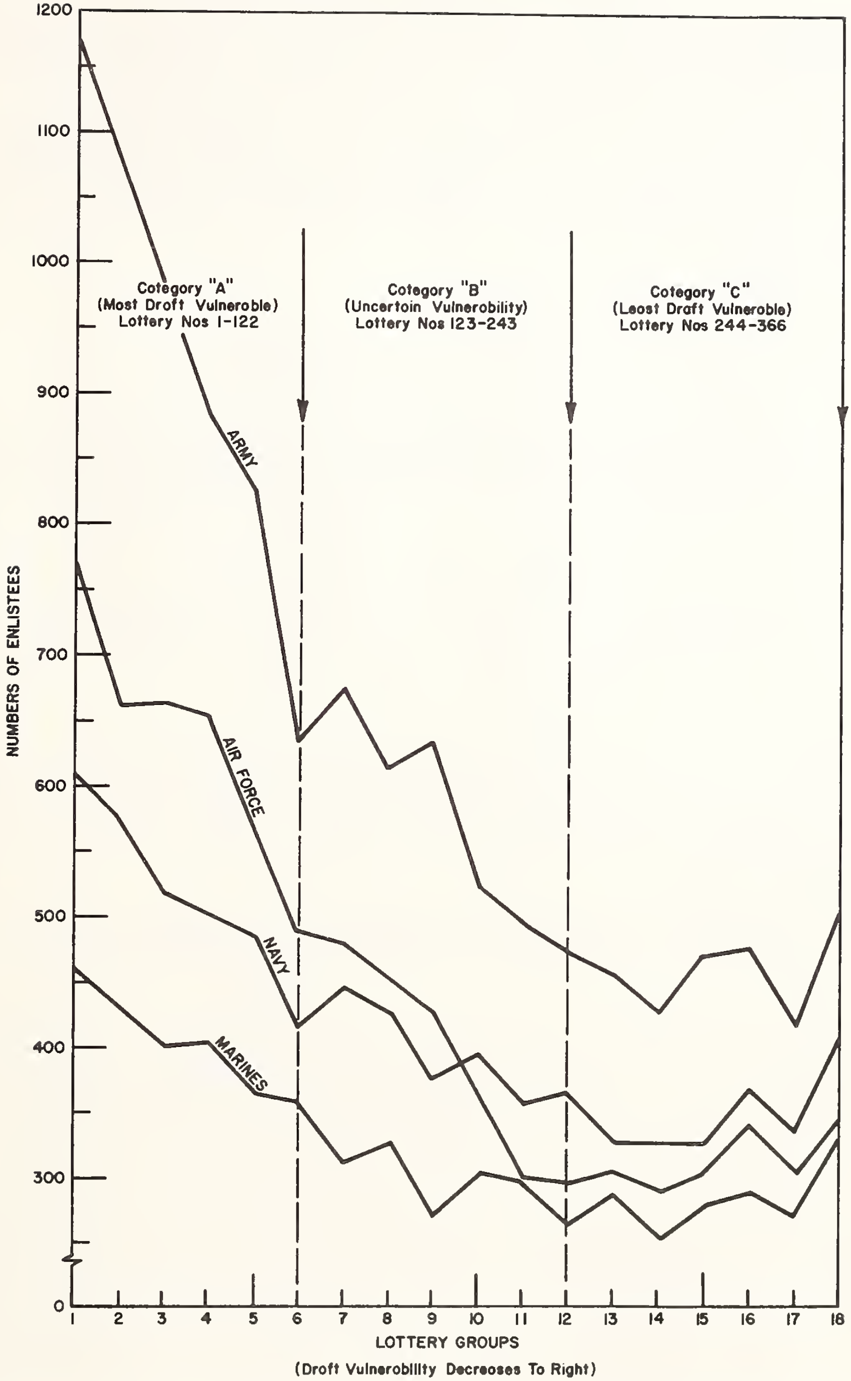


FIGURE 2. Draft lottery groups (draft vulnerability) vs numbers of enlistees/group (birth group 1951).

were in 1968 (due to subsequent introduction of the draft lottery). However, since there is no readily available method of determining a factor for Perceived Future Draft Vulnerability, the 1968 DOD attitude survey data were utilized in order to present a "pessimistic" estimate of true volunteers from the 1952–53 birth group. An "optimistic" estimate was also made based on the premise that 100 percent of this group are true volunteers.

These estimates are*

$$V_3 = \text{Number of enlistees in birth group 1951–53 (through June)} = 23,374$$

and

$$V_4 = \text{Number of enlistees in birth group 1952 and earlier (1952–53)} = 9,900.$$

In order to project the value of V_4 for the remainder of calendar year 1970, it was necessary to apply the adjustment factor (6/4).

Thus

$$V_K = V_3 + V_4(6/4)$$

$$V_K = 23,374 + 9,900\ (6/4) = 38,224,$$

where

$$V_K = \text{Optimistic estimate of total true volunteers from group not in lottery for 1970.}$$

Therefore, to estimate the total number of Navy volunteers (rounded down) of all birth groups for 1970:

$$V_O = 18,200 + 38,220 + 9,100 = 65,520,$$

where

$$V_o = \text{“Optimistic” estimate of total true volunteers for 1970.}$$

$$V_p = 51,760,$$

where

$$V_p = \text{“Pessimistic” estimate of total true volunteers for 1970.}$$

Table 1 presents "true" volunteer estimates for all Services, and was developed using methodology similar to that presented above.

TABLE 1. *Estimates of True Volunteers for All Services — 1970*

Estimates	Service			
	Army	Navy	Marines	Air Force
Optimistic	98,900	65,520	50,590	44,400
Pessimistic	77,820	51,760	42,940	37,850

Statistics utilized in this study indicate that the vast majority of service recruits fall into the age group of 17 to 22. Table 2 was formulated from data available from the Departments of Commerce and Labor and presents estimated increases in total male population between the ages of 17 and 22.

Table 3 presents a compilation of estimates of true volunteers for 1972 and 1973, based upon the increases in total male population reported above.

DETERMINATION OF ELASTICITY OF SUPPLY OF TRUE VOLUNTEERS WITH RESPECT TO MILITARY PAY

INTRODUCTION.

As stated previously, a primary objective of this study is to estimate the incremental MPN (Mili-

*Numerical values of V_3 and V_4 are taken directly from summary page of Annex A of Ref. [7] (for the Navy). Note that the value of V_4 covers the period 1 July to 31 October 1970 only.

TABLE II. *Increases in Total Male Population as a Percent of 1970 Total* [a]

Age group	Year	
	1972	1973
17–19	+5.7	+7.6
20–22	+3.9	+7.3

[a] Estimates based upon Table 1 "Population, total labor force, and labor force participation rates, by age and sex, actual 1960 and 1968 and projected 1975, 1980, and 1985," *U.S. Labor Force: Projections to 1985*, U.S. Department of Labor, February 1970, and Table 2 "Annual Projections of Population of the United States, by Single Years of Age and Sex: 1970 to 1985 with Quinquennial Extensions to 2020" *Population Estimates and Projections*, U.S. Department of Commerce, 6 August 1970.

TABLE 3. *Projected Estimates of True Volunteers for all Services, 1972–73*

Estimates	Army		Navy		Marines		Air Force	
	72	73	72	73	72	73	72	73
Optimistic..........................	104,000	106,000	69,000	70,000	53,000	54,000	47,000	48,000
Pessimistic.........................	82,000	84,000	54,000	56,000	28,000	28,000	32,000	33,000

tary Personnel, Navy) expenditures which would be necessary to sustain desired FY 1972–73 USN end strengths in a draft free environment.

Military Personnel, Navy expenditures in a given fiscal year may be subdivided into two parts: (1) the costs of the force in being, and (2) the costs of the new enlistees in that year. For present purposes, attention is focused primarily on the latter of these two factors, i.e., the total incremental expenditures necessary to insure that the supply of qualified volunteers meets or exceeds the demand for same in a given period. Assuming that the demand for qualified volunteers would exceed the supply in a draft-free environment (not an unfounded assumption), methods for increasing the supply of volunteers must be identified and corrective action taken lest actual end strengths fall short of those desired. Of course, there is an alternative method of insuring that supply meets or exceeds demand and that is to decrease demand. The most obvious single way to do so, short of reducing end strength requirements, is to improve the retention of personnel in the existing force. Although the consideration of retention rates is of prime importance in the intermediate or long-run, it is of lesser importance in the short-run (FY 1972, 1973), simply because the impact of retention-improving methods would probably be minimal in the early years of a draft-free environment. For this reason, improved retention rates have not been addressed explicitly in this study.

Returning to a discussion of the supply of volunteers, it is noted that supply is dependent upon many factors. Among these are military pay, educational benefits, opportunity for travel, recruiting methods, etc. In theory, appropriate changes in one or more of these factors could increase the supply of volunteers over current levels. However, only the impact of military pay on the supply of volunteers

is treated here. This paper will thus estimate the increase in the supply of volunteers, if any, which would result from a given increase in military pay.

The methodology employed to accomplish this task is one which has been applied fairly extensively in the 1964 and 1968 Presidential Commission studies on an all-volunteer force.* The methodology entails the determination of a quantitative relationship between the rate of supply of volunteers and relative military pay (relative to the income which a potential enlistee could earn in civilian life), based on historical data. Regression analysis is used to "fit" a curve to such data.

A hypothetical example of the results of such an analysis is shown in Figure 3.

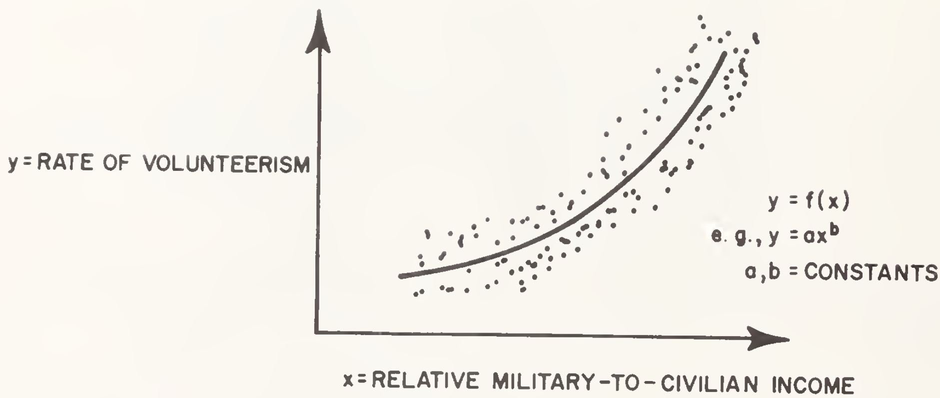


FIGURE 3. Hypothetical regression results.

The nomenclature for Figure 3 is as follows:

$$y = \text{true volunteer rate} = \frac{\text{no. of true volunteers}}{\text{total eligible population}}$$

$$x = \text{relative military-to-civilian income} = \frac{\text{military "pay"}}{\text{civilian "pay"}}$$

As noted in Figure 3, the regression analysis is used to aid in ascertaining whether or not there is a relationship between the independent variable x and the dependent variable y, and to estimate the form and parameters of the functional relationship, should such a relationship appear to exist. The functional form shown in Figure 3, $y = ax^b$, is given for illustrative purposes only. If a relationship or supply curve such as the one shown in Figure 3 can be estimated, then the *elasticity* of the supply of volunteers can also be estimated.

The elasticity of supply in this case is the ratio of the percentage change in rate of volunteers to the percentage change in relative military pay. (The "rate" of volunteers is the number of volunteers expressed as a percentage of "eligible" population).

$$\epsilon = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} = \frac{\Delta y}{\Delta x}\left(\frac{x}{y}\right),$$

where ϵ = supply elasticity,
y = true volunteer rate (as before),

*See, for example, References [1–3, 5, 6].

x = relative military-to-civilian income (as before),

Δy = the change in true volunteer rate, and

Δx = the change in relative military-to-civilian income.

The elasticity of true volunteer rate with respect to income may be estimated if the functional form and parameters of the supply curve (y vs x) is known. For example, if the functional form of the supply curve was

$$y = ax^b,$$

where a and b are constants to be estimated by regression techniques, then the elasticity, ϵ, can be determined from the elasticity relationship*

$$\epsilon = \frac{dy}{dx}\left(\frac{x}{y}\right) = abx^{b-1}\left(\frac{x}{y}\right) = \frac{b(ax^b)}{y} = b.$$

Thus, for this particular functional form, the elasticity is constant and is equal to b, a coefficient estimated using regression analysis.

Other functional forms for supply curve have varying elasticity. Consider, for example, the following illustrative relationship:

$$y = a + bx$$

(a, b, x, y as defined previously)

*The change to dy/dx from the $\frac{\Delta y}{\Delta x}$ used previously is made assuming that the changes Δy and Δx are small.

The elasticity of supply for this linear supply curve can be shown to be

$$\epsilon = \frac{y-a}{y}.$$

That is, the elasticity of supply (true volunteer rate) is a function of the current supply levels.

If elasticity can be estimated, then the pay increases required to increase the supply of volunteers can also be estimated.

The determination of statistical (based on historical data) supply curves and supply elasticities for for Navy (USN) and the Army are considered in this section. Army volunteerism and required accessions must be addressed in this study, based on the assumption that Army requirements will be the driving force in determination of new military pay scales in a draft-free environment. Thus, required pay increases to maintain Army end strengths in FY 1972, 1973 must be estimated. The effects of these pay increases on USN personnel budget, force strength, and quality can then be determined.

Similar estimates will be made for the Navy alone, to indicate USN pay requirements, assuming that Army requirements have no effect on the Navy. This could conceivably be the case if different pay increases are enacted for each Service—the Army presumably getting the largest increase based on their requirements. (This, however, would tend to drive some potential Navy volunteers into the Army.)

The implications of the results of this section on the quantity and quality of USN "true" volunteers are discussed in the final section of this report.

Input Data for Regression Analysis. In order to estimate a relationship between true* volunteer

*It is necessary to stipulate "true" volunteers when discussing enlistments in a draft environment.

rate and relative military-to-civilian pay, it was necessary to observe true volunteer rates at different relative pay levels. Since military pay is fixed for first-term enlistees, true volunteer rate could have been observed as a function of civilian pay by geographic region, age, education, or any other attribute of enlistees which results in civilian pay differentials. This study chose to utilize true volunteer rates for calendar year 1970 as a function of age and geographic region, as shown in Table 4.

TABLE 4. *Categories for True Volunteer Rate*

Age group	Geographic region
17–19	Northeast North Central South West
20–21	Northeast North Central South West

Enlistees in the age group 22–26 were aggregated into the 20–21 group, since the number of 22–26 year old enlistees was relatively small. Thus, each supply curve was estimated from eight data points; two age groups and four regions for each age group.

The determination of true volunteer rates and relative military-to-civilian pay factors for the eight data categories entailed estimation of the following quantities:

(1) Military "pay,"

(2) Civilian "pay" by age and region,

(3) The actual supply of true volunteers by age and region, and

(4) The size of the "eligible" population by age and region.

Each of these factors will be discussed in turn. The regression variables, repeated here for convenience are

$$x = \text{relative military-to-civilian pay} = \frac{\text{military pay}}{\text{civilian pay}}$$

and

$$y = \text{true volunteer rate} = \frac{\text{actual supply of true volunteers}}{\text{size of eligible population}}.$$

(1) *Military "Pay."* Military pay was defined as the annual base pay for enlisted grade E–3, or approximately $2,000 per annum. (Enlistees generally reach E–3 within a year after enlistment). Many variations on military pay were possible; e.g., these include allowances and imputed value of travel, education, and retirement benefits; determine the discounted value of "pay" over an *n*-period planning horizon, and so forth. However, cursory analysis indicated that all of these definitions would yield similar regression results.

(2) *Civilian "Pay."* It was attempted to obtain civilian income data by region, age, and attained educational level; however, this data were not available at the time of the study. The 1970 census data had not been processed and as of this writing are still not available. It was, therefore, necessary to use unpublished Bureau of Census data for 1969 which were consistent with Current Population Reports series. These data are shown in Tables 5 and 6.

TABLE 5. *Median Annual Income by Age for the United States (1969)*

(Source: Bureau of Census unpublished data consistent with Current Population Reports Series)

Age (yr)	Median annual income (dollars)
18–19	$3,198
20–21	$4,723

TABLE 6. *Median Annual Income for Males by Region (14 yr and over)*

(Source: Bureau of Census Report P–60, No. 66, Dec. 23, 1969, Table 52)

Region	Median annual income (dollars)
Northeast	$7,848
North Central	8,163
South	6,795
West	8,907
United States	7,814

A civilian income factor was then defined as follows:

$$C_{i_j} = A_i \left(\frac{R_j}{R}\right),$$

Where C_{i_j} = civilian income factor for age group i and region j,

A_i = median annual income for age group i (Table 5),

R_j = median annual income for males in region j (Table 6), and

R = median annual income for United States males (Table 6).

The values obtained in this manner were then divided into military pay ($2,000), resulting in the age/regional military-to-civilian pay factors shown in Table 7.

TABLE 7. *Age/Regional Relative Income Factors for Regression Analysis*

Age (yr)	Region	Relative income factor
17–19	Northeast	0.63
	North Central	0.60
	South	0.72
	West	0.55
20–21	Northeast	0.42
	North Central	0.41
	South	0.49
	West	0.37

Note that no claim is made that these factors represent actual age/regional median or mean relative incomes.

(3) *Actual Supply of True Volunteers.* There are two problems inherent in estimating the actual or true supply of true volunteers. The first is to identify enlistees as either draft-motivated or true volunteers. This was accomplished using draft lottery data, as indicated previously in this report. Secondly,

the number of enlistees may not be indicative of the total supply of potential enlistees if the supply of potential enlistees exceeds demand.

In order to separate supply from demand, it was necessary to identify those categories of true volunteers where demand was fairly certain to exceed supply. The actual number of true volunteers in such categories would then be indicative of true supply.

The method used, as in previous study efforts,* was to categorize true volunteers by mental ability, as measured by AFQT (Armed Forces Qualification Test) scores. The AFQT test is administered to prospective enlistees in order to ascertain mental fitness. The probability of acceptance of a potential enlistee is a function of his AFQT score, among other factors. AFQT scores are categorized as shown below (in descending order of mental ability):

AFQT category	Percentile range
I	92–100
II	64–91
IIIa	48–63
IIIb	31–47
IV	10–30
V	0–9

Category V individuals are generally not accepted. Category IV personnel are classified as "untrainable," and are generally considered least desirable for military service. However, each Service currently has a quota for Category IV enlistees under DOD "Project 100,000," which stipulates that 100,000 Category IV enlistees (total) must be accepted by the Armed Services annually.

Enlistees in Categories III, II, and I become more desirable (given physical fitness) as AFQT score increases. It is safe to assume that all Category I and II individuals who apply are accepted, i.e., the number of enlistees in these categories is the true supply. Category III is the gray area. Both analysis and conversations with BUPERS personnel indicate that some Category III individuals are turned away, i.e., the supply of these types of individuals exceeds demand. Further analysis has indicated that it is sound to assume that all or most Category IIIa individuals are accepted. Thus, it was assumed that the number of true volunteers in AFQT Categories I, II, and IIIa represented true supply. Only these true volunteer levels were used in analysis.

Data on USN enlistees for January–October 1970 by lottery category, age, region and AFQT category was obtained from BUPERS. Similar data was not available for the Army, so the assumption was made that the ratio of Army true volunteers in AFQT Categories I, II, and IIIa to total volunteers was identical to the Navy. These percentages are shown in Table 8.

(4) *The size of the "eligible" population by age and region.* An individual was defined as eligible if he would score in AFQT Categories I, II, or IIIa.† Since the *rate* of true volunteers was defined as the ratio of the number of true volunteers by age and region to the number of eligibles in that same age

*See References [1–3, 5, 6].

† Physical fitness was not considered, since available data indicated that physical reject rates were similar for the defined age groups and geographic regions. Thus, their inclusion in the analysis would have had little or no effect on the estimated supply curves.

TABLE 8. *Percentage of True Volunteers (USN Lottery Category "C" Enlistees; Jan–Oct. 1970) In AFQT Categories I, II, IIIa by Age and Region*

(Source: BUPERS Special Enlisted Master Tape Readout)

Age (yr)	Region	Percent of true volunteers (USN-lottery Category "C") in AFQT Categories I, II, or IIIa
17–19	Northeast	65.7
	North Central	68.8
	South	63.1
	West	77.1
20–21	Northeast	69.5
	North Central	71.3
	South	58.7
	West	73.1

group and region, it was necessary to determine estimates of this latter factor. To accomplish this task, the following three kinds of data were utilized:

(a) The total male population in the 17–19 and 20–21 age groups by region,

(b) The educational distribution of males by age and region, and

(c) The relationship between education level and AFQT category by region.

The total regional male population estimates used (1970) are shown in Table 9. (The data and methodology upon which these estimates are based is shown in Annex B of Ref. [7].)

TABLE 9. *Estimated Total Male Population by Age and Region (1970)*

Region	Total male population 17–19 yr	Total male population 20–21 yr
Northeast	883,000	534,000
North Central	1,738,000	1,010,000
South	2,321,000	1,488,000
West	718,000	538,000

Recent educational data by age and region was not available. The available data was in two forms; (1) educational distribution by age and region for 1960 and, (2) educational distribution by age for the United States (1969). The 1960 census data were outdated and was therefore not used. These data did indicate, however, that the educational distributions for the Northeast, North Central, and West regions were similar; educational attainment in the South being somewhat lower. Thus, the decision was made to assume that educational attainment in the Northeast, North Central, and West in 1970 is identical to United States educational attainment in 1969. The educational levels in the South were varied parametrically in the regression analysis and, fortunately, regression results were insensitive to these variations.

Data relating AFQT category and educational level were obtained from a 1964 study involving 50 percent of the 1963–1964 inductees [4]. This data was combined with educational and population data to yield estimates of percentages and numbers of eligibles (AFQT I, II, or IIIa) by age and region, as shown in Table 10.

TABLE 10.[a] *Estimated Percentages and Numbers of Eligible Males (AFQT I, II, IIIa) by Age and Region*
(Source: See [7, Annex B] for derivation)

Region	Population eligible (%)		Number of eligibles	
	17–19 yr	20–21 yr	17–19 yr	20–21 yr
Northeast	39.58	60.48	570,000	460,000
North Central	43.42	65.63	720,000	580,000
South	30.82	54.39	570,000	540,000
West	51.29	71.02	510,000	370,000

[a] Figures for the South are based on the assumption that educational attainment for the South was identical to the overall United States educational attainment. Since this is biased on the high side, the number of South eligibles was varied parametrically downward in the analysis, with insignificant effect on results.

The preceding data were combined with the true volunteer estimates (Table 1) to yield the final data for regression analysis. This analysis follows:

Regression Results. Data on eligible population and relative income factor by age and region were combined with Army and Navy true volunteer estimates for calendar 1970 (Table 1) yielding the regression results shown in Figure 4 (U.S. Navy) and Figure 5 (U.S. Army). Data sheets for these regression analyses may be found in Ref. [7, Annex C].

The supply curves shown in Figures 4 and 5 were fitted to the data using the method or ordinary least squares. A first step in this analysis involved the choice of a model, e.g., linear concave, convex, etc. for each particular model. Ordinary least squares will yield a "best-fitting" line for each model, but this does not imply that any of these will fit the data well.

The more important forms of possible models can be classified as constant elasticity models or decreasing elasticity. The form of the constant elasticity model is

$$y = ax^b$$

$$\epsilon = \text{elasticity} = b.$$

Table 11 shows some decreasing elasticity models.*

TABLE 11. *Some Decreasing Elasticity Models*

Model	Functional form[a]	Elasticity (ϵ)
Complement	$(1-y)=ax^b$	$b\left(\frac{1-y}{y}\right)$
Gompertz	$y=e^{-ax^{-b}}$	$-b\ln y$
Logistic	$\frac{y}{1-y}=e^{ax+b}$	$bx(1-y)$
Log logistic	$\frac{y}{1-y}=ax^b$	$b(1-y)$

[a] e = the base of the natural logarithms.

Economic theory suggests that real supply curves are characterized by decreasing elasticity, i.e., elasticity of supply decreases as actual supply increases.

*See Ref. [3] for a more comprehensive discussion of these models.

Each of the five models was fitted to USN (Figure 4) and USA (Figure 5) true volunteer data, respectively. All of the models fit the USN data about equally well as indicated by various statistical measures such as correlation coefficient, t-value and residuals checks. Moreover, each model fitted the data fairly well in an absolute sense (correlation coefficient 0.80–0.85). The same kinds of comments apply to the analysis of Army data. Thus, although there appears to be a discernable relationship between true volunteer rate (y) and relative military-to-civilian pay, it is impossible to distinguish between constant and varying elasticity models on the basis of the data. However, one consolation was the fact that estimated supply elasticities were comparable, regardless of the form of model chosen.*

The constant elasticity model was chosen for use in this study, thus the least squares curves shown in Figures 4 and 5 are of the functional form

$$y = ax^b,$$

where estimated elasticity is constant and is equal to the regression coefficient b. The choice of this model was based neither on theory nor statistical analysis of data. It is merely the case that the majority of the models yielded comparable results, and that the constant elasticity model lends itself to discussion and interpretation. A brief discussion of regression results follows.

1. *USN Regression Results.* The estimated constant elasticity USN supply curve is: $y = 2.75x^{1.6}$ indicating an estimated supply elasticity of 1.6. The measures of goodness-of-fit are:

R^2 (proportion of sum of squares reduced) $= 0.77$,

R (correlation coefficient) $= \sqrt{0.77} = 0.87$, and

t-value (for b) $= 4.432$.

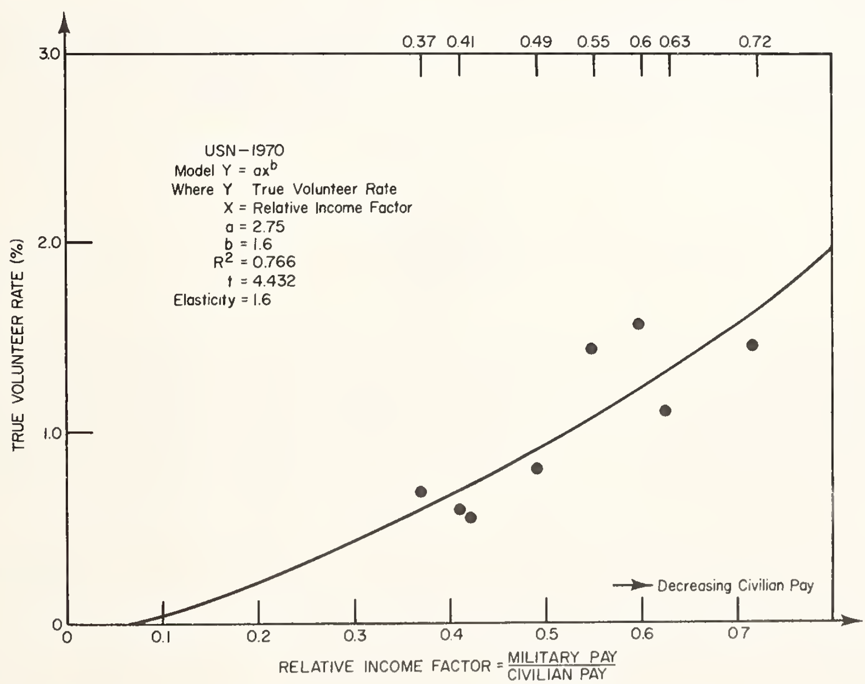


FIGURE 4. Estimated U.S. supply curve (true volunteers-mental groups I, II, IIIA).

*An exception is the complement model, which consistently yielded elasticity estimates which were 50–100 percent greater than elasticities derived from the other models.

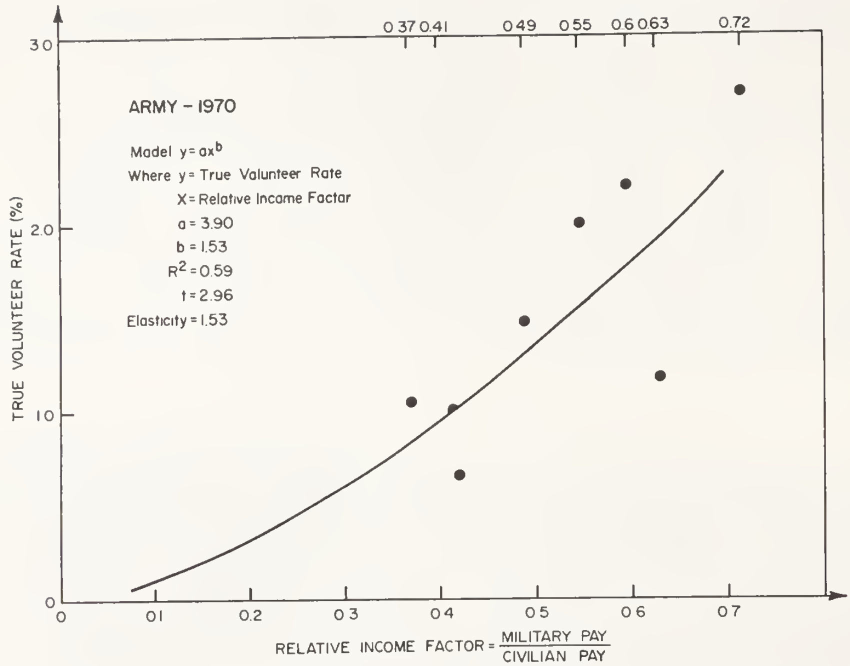


FIGURE 5. Estimated supply curve for U.S. Army (true volunteer in mental groups I, II, IIIA).

Moreover, an examination of residuals (deviations of observed true volunteer rate from predicted rates using the regression curve) revealed no anomalies.

2. *USA Regression Results.* Elasticity of the estimated Army supply curve was 1.53. The fit of the regression curve for the Army (Figure 5) was not as good as the fit of the Navy supply curve (Figure 6), but overall results can be classified as acceptable, as indicated below:

$$y = 3.9x^{1.53}$$
$$R^2 = 0.59$$
$$R = 0.78$$
$$t = 2.96.$$

Thus, on the basis of the preceding analyses, it was concluded that USN and USA true volunteer rates were elastic ($\epsilon \cong 1.5$), indicating that true volunteer rates for USN and USA, respectively, would increase 1.5 percent for every 1-percent increase in military base pay, grades E−1 through E−3.*

Evaluating Figure 6 would seem to indicate that 17–19 year olds have a greater propensity to volunteer than do 20 to 21 year olds. This alternative hypothesis was not tested because in the set of data the factor of age and income were statistically confounded. However, it would appear that if data had been available which would have allowed breaking out the volunteers by year of birth and correlating these data with the volunteer rate very little change would have been noted. There is a considerable body of evidence that lower income levels produce a greater volunteer rate and thus the authors felt that relating volunteer rate to the income factor was a better explanatory variable than age.

*A 1 percent increase in military pay is the same as a 1 percent increase in relative military-to-civilian pay, for a given civilian pay level.

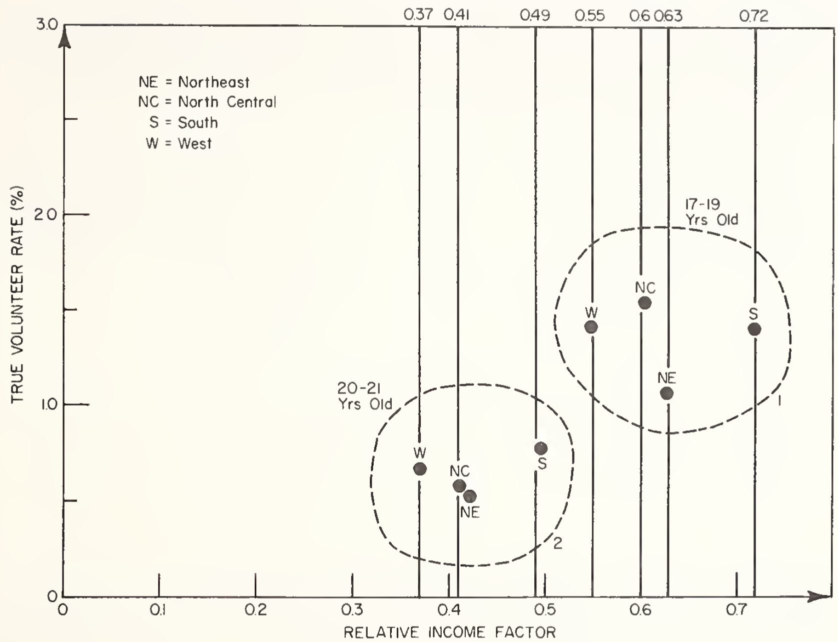


FIGURE 6. U.S. Navy results—sensitivity analysis.

It is possible at this point to move on and discuss the implications of these results on MPN budget and USN accessions and quality profile. However, some fairly extensive sensitivity analyses have indicated that results (supply elasticities and goodness-of-fit of regression models) were sensitive to a number of the assumptions which were made in the course of analysis. It is worthwhile at this point to discuss these sensitivity analyses briefly.

Sensitivity Analyses on Regression Models. The purposes of regression analysis were twofold:

(1) To ascertain if a relationship exists between true volunteer rate and relative military-to-civilian pay, and

(2) To estimate supply elasticities should such a relationship exist. An investigation was made of those assumptions which, when varied, would affect either (1) or (2) above.

To facilitate the ensuing discussion, it is useful to refer to Figure 6, which is a repeat of the USN data shown in Figure 4; however, Figure 6 shows that the total data points are comprised of two data subsets: one for the 17–19 year age group (group 1), the other for the 20–21 year age group (group 2). It is apparent from Figure 6 that any change which would shift one of these subsets of points up or down relative to the other could alter results substantially. Accordingly, certain potential causes of such shifts were identified and investigated. The more important potential causes are discussed below.

(1) The actual true volunteer rate for 17–19 year olds may be less than indicated. Since individuals in this group do not have lottery numbers, it is not possible to identify true volunteers by lottery analysis. Since true volunteer rates shown in Figure 6 (and in Figure 5—Army) were derived by assuming that *all* 17–19 year old enlistees in calendar 1970 are true volunteers, it is appropriate to examine the effects of lowering true volunteer rates for this age group. It is obvious from Figure 6 that the overall effect of such a change would be to decrease the estimated elasticity.

Accordingly, true volunteer rates for 17–19 year-old enlistees for USN and USA, respectively, were decreased by 25 percent. The 25-percent figure was based on the assumption that the percentage

of draft-motivated enlistees in the 17–19 age group falls somewhere between 0 and 35 percent (a 1968 DOD survey indicated about 35 percent draft-motivated for 17–18 age group—Army and Navy, respectively. See page 10). The effects of this variation on regression results are noted below.

USN	*USA*
Constant elasticity model	Constant elasticity model
Elasticity estimate = 1.02	Elasticity estimate = 0.94
$R^2 = 0.67$	$R^2 = 0.36$
$R = 0.82$	$R = 0.6$
t-value = 3.52	t-value = 1.837

These results indicate a 30-percent decrease in estimated supply elasticity and a degradation in the fit of the models, particularly for the Army. The fit of both models is still considered acceptable, however.

(2) The size of the eligible population for the 20–21 age group may be relatively high. All United States males aged 20–21 with predicted AFQT I, II, or IIIa mental ability were assumed to be eligible. However, this assumption ignores the fact that some non-negligible proportion of this age group population is comprised of individuals who have already completed or are currently completing military service. Such individuals would be considered ineligible. The same kind of statement applies to the 17–19 age group, but probably to a lesser degree. Thus, the net effect of refining the size of the eligible populations may be to shift the 20–21 true volunteer rate upward relative to the 17–19 rate (refer to Figure 6). Accordingly, the eligible population for the 20–21 age group was decreased by 15 percent. Results of this change are shown below.

USN	*USA*
Constant elasticity model	Constant elasticity model
Elasticity estimate = 1.27	Elasticity estimate = 1.18
$R^2 = 0.73$	$R^2 = 0.48$
$R = 0.85$	$R = 0.69$
t-value = 4.0	t-value = 2.376

Results indicate approximately a 20-percent decrease in estimated elasticity. Fit of the models does not change significantly.

The combination of a 10-percent decrease in 17–19 age group true volunteer levels and 10-percent decrease in 20–21 age group population yielded similar results—elasticities about 1.10 and good correlation. (It was felt that decreasing 17–19 age group true volunteer levels by 25 percent and decreasing 20–21 age group eligible population by 15 percent simultaneously was an unreasonably conservative procedure).

Results were found to be insensitive to the following inputs:

(1) Regional age distributions.

(2) Regional education profiles (e.g., 1960 census data or 1969 data for males 25 and over, or 1969 United States education distribution for 17–19 and 20–21 age groups, respectively, applied to all regions).

(3) (1) and (2) together.

Summary. True volunteer rates and income are correlated. Estimates of elasticity of true volunteer rate with income appear to range between 0.9 and 1.5, with a most likely value of 1.2.

THE SHORT TERM IMPLICATIONS OF AN ALL-VOLUNTEER NAVY

The impact of military pay on USN FY 1972, 1973 accessions, both in terms of quantity and quality, are examined in this section. Military pay will be measured in terms of percentage increase in base pay and also in terms of incremental MPN expenditures. Quantity is simply the number of enlistees which can be expected in a draft-free environment. The measure of an enlistee's "quality" will be his AFQT category, since this is the most commonly accepted quality measure.

There are several problems inherent in a discussion such as this. First, base pay increases for enlisted grades E–1 through E–3 cannot realistically be treated without considering also the effects of such increases on E–4 through E–9 pay scales. For the most part, this is due to the fact that an acceptable "payline" must be maintained. A payline can be viewed simply as the relationship between base pay and enlisted grade, as shown in Figure 7. Note that the payline is an increasing function of enlisted grade. If, however, the base pay for grades E–1 through E–3 were to be raised 40 percent (for example) with no attendant increase in base pay for grades E–4 through E–9, the positive slope (increasing function) of the payline would be violated at grade E–4, as shown in Figure 8.

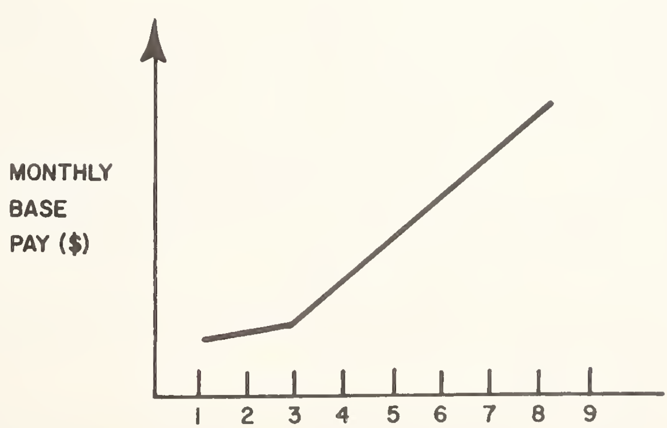


FIGURE 7. An enlisted payline, 2 years service or less.

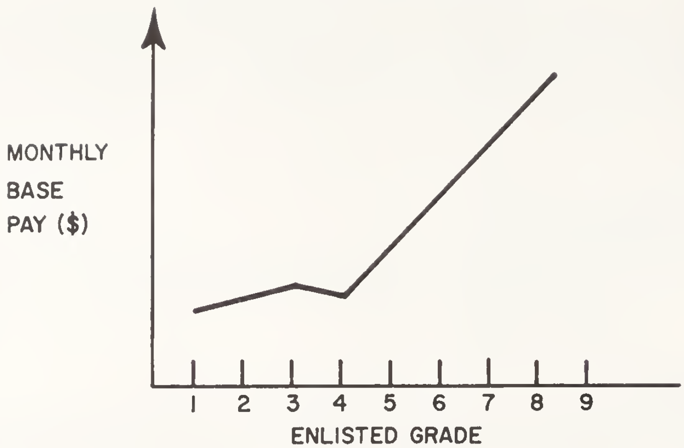


FIGURE 8. A hypothetical payline.

Thus, if first term (E–1 through E–3) pay increases were to exceed, say 20 percent, the payline becomes an extremely important factor in estimating the total incremental expenditures for the "military pay increase." The problem is not so much estimation of MPN expenditures for a given payline and force profile, but rather it is the formulation of the payline itself.

Another fundamental consideration is the determination of the planning horizon for budgeting purposes. If budget requirements are formulated based on long-term accession requirements, and budget appropriations are equal to requirements, then it is possible that short-term accession and quality levels will fall below those desired. If, on the other hand, budget appropriations are adequate to maintain desired accession and quality levels in the short-term, it is possible that the supply of desirable volunteers would unnecessarily exceed demand in the long-term. This is because long-term improvements are intended to make the service more desirable and thus pay would become less important.

Finally, for purposes of this discussion at least, there is the distinct possibility that a "lump" annual sum will be appropriated to the Navy, and the Navy will retain the responsibility for determining optimal allocation of this money among pay increases, enlistment bonuses, recruiting, and so forth.

Yet, all these considerations notwithstanding, there remains the fact that future budget requirements must be formulated. To do so, the short-term/long-term problem must first be resolved. The approach taken here involves estimating the incremental monetary requirements necessary to insure maintenance of desired end strengths and quality levels in the short-term. Equally important however, the short-term implications of "insufficient" budget appropriations must be examined, so that the Navy can develop viable contingency plans for the short-term, if not also for the long-term.

This section includes discussions of the following relevant considerations:

(1) The quantity and quality of all enlistees and of true volunteers in calendar year 1970.

(2) A "most likely" all-volunteer force for calendar year 1970, assuming present pay levels.

(3) An all-volunteer USN in FY 1972, 1973 assuming that all services receive identical pay increases, based on Army requirements.

(4) USN budget requirements for FY 1972, 1973 in the absence of Army considerations.

(5) The effects of "insufficient" budget appropriations on the supply and quality of volunteers in FY 1972, 1973.

The quantity and quality of total enlistees vs true volunteers for calendar year 1970. As indicated in Part I of this report, the number of USN true volunteers for calendar year 1970 is estimated to be between 51,760 and 65,520, with the "most likely" estimate closer to 65,000 than 51,000. These figures indicate that true volunteers represent from 51 to 71 percent of total USN accessions in calendar year 1970.

The quality of enlistees in calendar year 1970 can be illustrated both by educational distribution and AFQT distribution. Education is discussed first.

Statistical data was made available by the Bureau of Naval Personnel* listing the educational attainment and racial characteristics of first term enlistees entering the Navy between January and September 1970. These data were organized and displayed for three basic groups according to previously defined categories for those in the 1970 lottery:

Category A—Certain to be Drafted,

Category B—Uncertain, and

Category C—Certain not to be Drafted.

Table 12 presents a compilation of educational data for those enlistees born between 1944 and 1950 (those in the lottery), by draft lottery category and racial characteristic.

TABLE 12. *Educational Attainment for Those Enlistees with Birth Years 1944–50 by Draft Lottery Category and Race*

Draft category and race \ Educational attainment	Years of education				
	0–8	9–11	12	13+	Total (%)
CAT "A"					
White	0.7	7.8	56.2	35.3	100
Non-white	0.4	13.2	63.9	22.5	100
All	0.7	8.4	57.0	33.9	100
CAT "B"					
White	0.8	9.0	57.2	33.0	100
Non-white	0.6	16.9	63.1	19.4	100
All	0.7	9.9	58.0	31.4	100
CAT "C"					
White	1.0	11.8	60.7	26.5	100
Non-white	1.3	17.2	64.5	17.0	100
All	1.1	12.8	61.4	24.7	100

Table 13 presents educational data for those enlistees with birthdate between 1950 and 1953 by race only since this group is not in the lottery system.

TABLE 13. *Educational Attainment for Those Enlistees with Birth Years 1951–53 by Race*

Racial character \ Educational attainment	Years of education				
	0–8	9–11	12	13+	Total (%)
White	3.3	33.4	62.0	1.3	100
Non-white	2.3	42.0	54.4	1.3	100

* Data Supplied by Pers–N211 included Educational Attainment of Enlistees who entered between January and September 1970, by race, by draft lottery category, age and geographic region of residence at time of entrance.

One overriding characteristic is evident from an inspection of the tables presented above. That is that the educational attainment of the most draft vulnerable groups (those in lottery category "A") is higher than that of least draft vulnerable group (those in lottery category "C"). The differences in educational attainment between white and non-white enlistees are also notable. Table 14 presents educational data for the United States population for comparison purposes.

TABLE 14. *Educational Attainment of Males (all races) Between Ages of 16 and 29 for the United States*

Age \ Education attainment	Percent distribution (all males)				
	0–8	9–11	12	13+	Totals (%)
16–17	12.3	86.5	1.0	0.2	100
18–19	6.7	38.6	41.9	12.8	100
20–21	7.0	14.6	33.3	45.1	100
22–24	8.0	14.2	40.0	37.8	100
25–29	9.5	14.9	39.9	35.7	100

Mental quality of the Armed Forces is also commonly indicated in one of two other ways; AFQT distribution or percentage of enlistees in AFQT categories I (highest achievers), II, or IIIa. Between 1960 and 1968 the percentage of USN enlistees in AFQT categories I, II, or IIIa has steadily increased from 60 percent to approximately 80 percent of total USN enlistees. Table 15 shows a comparison of mental abilities of true volunteers and draft-motivated enlistees based on this latter measure (for enlistees with birth years 1944–50).

TABLE 15. *Mental Fitness for Enlistees with 1944–50 Birth Years*

Lottery category	Type of enlistee	Percent in AFQT I, II, IIIa	Percent in AFQT IV, (untrainable)
A	Draft motivated and true volunteer	78	13
B	True volunteer only	66	20
	Draft-motivated only	83	9
All	Draft-motivated and true volunteer	75	15

As Table 15 shows, there are marked differences in the mental characteristics of true volunteers, draft motivated enlistees, and total enlistees (true volunteers and draft-motivated enlistees) for those enlistees with birth years 1944–50.

The mental characteristics of enlistees with birth years 1951–53 are shown in Table 16. Since some unknown proportion of this group may be draft-motivated (presumably with higher AFQT capability than true volunteers in the same group), it is reasonable to assume that mental characteristics of true volunteers in this age group are not quite as good as those indicated in Table 16. Mental characteristics of true volunteers with birth years 1944–50 are included for comparison, as are the characteristics for total USN accessions (all ages).

TABLE 16. *Mental Characteristics of Selected Enlistee Groups (Jan–Oct 1970)*

Type of enlistee	Year of birth	Percent in AFQT category	
		I, II, IIIa	IV (untrainable)
True volunteer and some draft-motivated	1951–1953	68	17
True volunteer	1944–1950	66	20
Total USN	1944–1953	71	16

The differences in mental ability between true volunteers and total enlistees as depicted in Tables 15 and 16 are notable. Yet, the discussion can be pursued still further, i.e., the following question can be addressed: What would be the mental profile of an all-volunteer Navy in 1970, assuming current pay levels?

A "Most Likely" All-Volunteer USN for Calendar Year 1970. All available data suggest that the number of enlistees in AFQT categories I, II, and IIIa represents the true supply of these kinds of enlistees, i.e., the Navy does not currently reject significant numbers of physically fit applicants of this mental calibre. However, there is substantial evidence that relatively large (thousands) numbers of applicants in AFQT Category IIIb are being rejected by the Navy this year. If this is true then the Navy could achieve higher accessions in 1970, without any pay increase, by accepting more IIIb applicants. The size of the excess supply of IIIb true volunteers, if any, can be estimated as follows.

The estimated numbers and percentages of USN true volunteers in AFQT Categories IIIa and IIIb for calendar year 1970 are shown in Table 17. These estimates are based on the optimistic true volunteer numbers shown in Table 1 (Part 1).

TABLE 17. *Estimated Numbers and Percentages of 1970 USN True Volunteers in AFQT Categories IIIa and IIIb*

AFQT category	Est. number of true volunteers	True volunteers (percent)
IIIa	19,100	29.2
IIIb	8,520	13.1

It is notable here that the estimated number of true volunteer enlistees in AFQT Category IIIa exceeds the number in Category IIIb by approximately 10,500. This information becomes useful when compared with the distribution of mental ability of 17–21 age group males for the total United States shown below.

AFQT category	United States males (17–21 yr) in AFQT category (estimated percent)
IIIa	18
IIIb	20

Thus, although the number of AFQT IIIb calibre males in the total United States population exceeds the number of AFQT IIIa calibre males, the USN Category IIIa enlistments far exceed those in Category IIIb.

Assuming that the rate of supply of true volunteer applicants of Category IIIb mental calibre is equal to the rate for applicants of Category IIIa calibre, then it appears that the Navy has rejected a minimum of 10,500–13,000 Category IIIb true volunteers this year. Moreover, the supply rate for IIIb true volunteers may actually exceed the supply rate of IIIa true volunteers. Thus it is entirely plausible to assume that the actual total supply of AFQT Category IIIb exceeds, say 21,000 for 1970, 8,500 of whom were accepted for service.

Based on optimistic estimates,* 65,000 of the 85,000 USN accessions in 1970 are true volunteers. Thus, in this case, 20,000 additional volunteers would be needed in the absence of a draft. Conceivably, this entire 20,000-man deficit could be eliminated in 1970 with no pay increase, if greater numbers of AFQT Category IIIb (and possibly IV) were accepted. For illustrative purposes, the plausible assumption has been made that an additional 20,000 true volunteers in AFQT Category IIIb exist at this time, with no pay increase. If this were the case, then the quality profile of USN 1970 accessions would change markedly, as indicated below.

Force	Accessions in AFQT I, II, or IIIa (percent)
Actual calendar 1970 USN	71
Estimated all-volunteer 1970 USN in absence of pay increase	53

*Recall that the "optimistic" estimates are based on the assumption that all USN 1970 enlistees with birth years 1951–1953 are true volunteers.

The total AFQT distribution for both the actual and estimated all-volunteer Navy accessions are shown in Figure 9. The horizontally striped bars of Figure 9 show the distribution for actual 1970 USN enlistees; whereas the vertically striped bars represent the estimated mental distribution for a hypothetical 1970 USN all-volunteer force in a draft-free, no-pay-increase environment. Total accession numbers are the same for both forces in Figure 9: i.e., the number of accessions (total) is 85,000 in both cases.

Thus, it appears that the Navy could conceivably have fulfilled accession requirements this year with no draft and no pay increase. However, as noted above, the overall mental distribution of enlistees would be shifted downward.

The cost of maintaining accession rates and quality levels of 1970 in the absence of a draft would require some kind of pay increase. Assuming that the Navy desires to maintain present quality levels (71 percent in AFQT Categories I, II, or IIIa) implies that 60,500 true volunteers in AFQT Categories I, II, and IIIa are required, for an accession rate of 85,000. Since the current "optimistic" estimate of the number of true volunteers in this category is 44,500, this implies that 16,000 additional volunteers (a 36-percent increase) are required, at the least. Assuming an elasticity value of 1.2 implies that base first term pay must be increased by 30 percent to attract the additional 36 percent volunteers for AFQT Categories I, II, and IIIa.

A 30 percent increase in base pay for enlistees in grades E–1 through E–3 would require additional MPN expenditures of approximately $80–100 million annually, based on current force structure.*

*BUPERS Cost Model (BUCOMP) results indicate an additional MPN cost of $90.8M for FY 72 if E–1, E–3 pay is increased 30 percent.

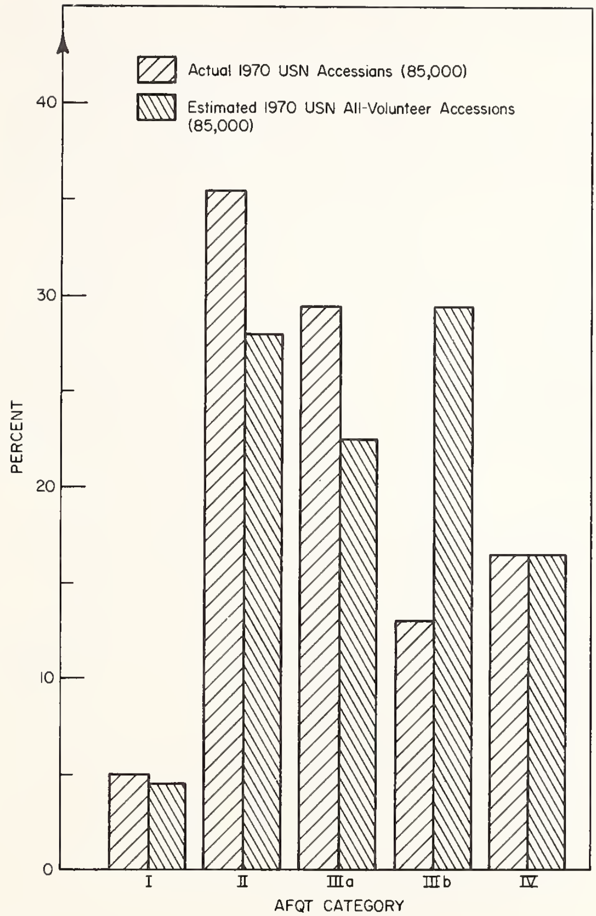


FIGURE 9. AFQT distributors (assuming 16 percent category IV's required per DOD "Project 100,000").

This figure is based on the assumption that pay scales for grades E–4 through E–9 remain constant. If, for the sake of illustration, E–4 through E–9 pay were increased 10 percent in addition to the 30 percent E–1 through E–3 increase, the total estimated incremental expenditure becomes about $300 million annually. Note that these estimates are based on the assumption that all 1970 USN enlistees with birth years 1951–1953 are true volunteers. If only 65 percent of these enlistees are true volunteers (the pessimistic estimate per 1968 DOD survey), then the required increase in AFQT I, II, or IIIa accessions would be approximately 70 percent, resulting in a required pay increase for grades E–1 through E–3 of 58 percent (assuming an elasticity of 1.2). Total incremental MPN expenditures resulting from such an increase are estimated to be $175 million annually, assuming that E–4 through E–9 pay is held constant.

Estimated Incremental Expenditures for an All-Volunteer Navy—FY '72, '73. USN accession requirements for FY '72 and FY '73 are forecasted to be 85,000 and 135,000 respectively. Table 18 shows estimates of the base pay increases required to attain these accession levels and at the same time maintain 70 percent of total enlistments in AFQT Categories I, II, and IIIa (70 percent of calendar year 1970 enlistments are in AFQT Categories I, II, IIIa). The "most likely" cost estimates for this to occur are summarized in Table 19 (elasticity = 1.2).

The total MPN expenditures shown in Table 19 are subdivided into those expenditures for enlisted grades E–1 through E–3, and those for enlisted grades E–4 through E–9. MPN estimates for grades

TABLE 18. *Required Pay Increases and Incremental MPN Expenditures for Enlisted Grades E–1 Through E–3 Based on Navy Accession Levels for FY '72, '73*

(70 percent of USN Accessions in AFQT Categories I, II, IIIa)

Estimates	Total USN accessions required (thousands)	Total USN accessions required in AFQT I, II, IIIa (70%) (thousands)	Estimated actual accessions in AFQT I, II, IIIa if no pay increase (thousands)	Additional AFQT I, II, IIIa accessions required (thousands)	Additional accessions in AFQT I, II, IIIa required (percent)	Pay increase required (percent)			Incremental MPN expenditures required for grades E–1 thru E–3 (millions)		
						$\epsilon = 0.9$	$\epsilon = 1.2$	$\epsilon = 1.5$	$\epsilon = 0.9$	$\epsilon = 1.2$	$\epsilon = 1.5$
Optimistic	85	60	47	13	28	31	23	19	95	70	60
	105	73	47	26	55	61	46	37	185	140	110
	135	94	47	47	100	110	83	67	330	250	200
Pessimistic	85	60	37	23	62	69	51	41	210	155	125
	105	73	37	36	97	108	81	64	325	245	190
	135	94	37	57	154	171	128	103	515	385	310

E–1 through E–3 were derived using the BUPERS cost model and the Enlisted Personnel Strength Plan, 72"B" Mod. This model indicates additional MPN FY '72 expenditures of $30 million (for grades E–1 through E–3) for each 10-percent increase in first-term base pay.

MPN estimates for grades E–4 through E–9 are based on increasing the average pay scales for these grades by decreasing percentages with grade. For example, if first-term base is increased 20 percent, E–4 pay would increase by 10 percent, E–5 pay by 9 percent and so forth. (Annex E includes a more comprehensive discussion of this topic).

The foregoing cost estimates are based on Navy requirements, as though these would be the driving force in determining military pay requirements in a draft-free environment. In fact, however, since salaries are set DOD-wide, Army requirements appear to be the driving force. For example, it is estimated that more than 70 percent of Army accessions in FY '70 (including inductees) were draft-motivated. Thus the Army has the greatest gap to close in an all-volunteer environment, and its requirements may well be the deciding factor in determining new military pay scales.

TABLE 19. *Summary of Required Base Pay Increases and Incremental MPN Expenditures Based on Navy Accession Requirements for FY '72, '73*

Required number of Navy accessions	Increase in first term base pay (percent)	Incremental MPN expenditures for grades E–1 thru E–3 (millions-constant 1970 dollars)	Incremental MPN expenditures for grades E–4 thru E–9 (millions-constant 1970 dollars)	Total incremental MPN expenditures: grades E–1 thru E–4
85,000	23–51	70–150	130–280	200–430
105,000	46–81	140–240	250–430	390–670
135,000	83–128	250–380	440–680	690–1060

Tables 20 and 21 show the impact of the Army on USN accession levels and incremental budget expenditures, assuming that pay increases are sufficient to enable the Army to meet FY '72 accession requirements and maintain 60 percent enlistees in AFQT I, II, and IIIa (the estimated current level). Table 20 shows "most likely" required pay increases and the impact of these increases on Navy MPN expenditures (elasticity = 1.2). Table 21 shows the same information in more detail.

TABLE 20. *Summary of Required Base Pay Increases and Incremental (Navy) MPN Expenditures for FY '72, '73 Assuming That Pay Increases are Based on Army Requirements*

Required number of Army accessions	Increase in first term base pay (percent)	Incremental MPN expenditures for grades E–1 thru E–3 (millions-constant 1970 dollars)	Incremental MPN expenditures for grades E–4 thru E–9 (millions-constant 1970 dollars)	Total incremental MPN expend; grades E–1 thru E–4 (millions-constant 1970 dollars)
180	45–80	140–240	240–430	380–670
200	59–98	180–290	320–520	500–810
220	73–117	220–350	390–620	610–970

Navy MPN expenditures are compared in Table 22 for the two cases previously discussed:

(1) Military pay increases based on Navy requirements.

(2) Military pay increases based on Army requirements.

TABLE 21. *The Effect of FY '72, '73 Army Accession Requirements on USN Base Pay, Accession Levels and Incremental MPN Expenditures for Enlisted Grades E–1 Through E–3*

(60 percent of Army Accessions in AFQT Categories I, II, and IIIa)

Estimates	Total Army accessions required (thousands)	Total Army accessions required in AFQT I, II, IIIa (60%) (thousands)	Estimated number of Army accessions in AFQT I, II, IIIa if no pay increase (thousands)	Additional AFQT I, II, IIIa accessions required (Army) (thousands)	Increase in Army AFQT I, II, IIIa accessions (percent)	Estimated number of USN accessions in AFQT I, II, IIIa if no pay increase (thousands)	Estimated USN accessions in AFQT I, II, IIIa if pay increase required by Army (thousands)	USN accessions in AFQT I, II, IIIa for following USN accession levels (percent)			Base pay increase for following elasticities (percent)			Incremental MPN expenditures for grades E–1 thru E–3		
								85	105	135	$\epsilon=0.9$	$\epsilon=1.2$	$\epsilon=1.5$	$\epsilon=0.9$	$\epsilon=1.2$	$\epsilon=1.5$
Optimistic	180	108	70	38	54	47	72	84	68	53	60	45	36	180	135	110
	200	120	70	50	71	47	80	94	76	59	79	59	47	240	180	140
	220	132	70	62	88	47	88	100	84	65	98	73	59	295	220	180
Pessimistic	180	108	55	53	96	37	72	84	68	53	107	80	64	320	240	190
	200	120	55	65	118	37	80	94	76	59	130	98	79	390	295	240
	220	132	55	77	140	37	88	100	84	65	156	117	93	470	350	280

TABLE 22. *A Comparison of Incremental Navy MPN Expenditures for FY '72 or FY '73 All Volunteer Force*

Required USN accessions	Pay increase based on Navy requirements		Pay increase based on Army requirements (assuming 200,000 accessions)	
	First term base pay increase (percent)	Total incremental MPN expenditures (millions)	First term base pay increase (percent)	Total incremental MPN expenditures (millions)
85,000	23–51	$200–430	59–98	$500–810
105,000	46–81	390–670	59–98	500–810
135,000	83–128	690–1060	59–98	500–810

Note that Army requirements necessitate higher base pay increases than do the Navy's, except when Navy accession requirements are 135,000. In this case, Navy accession requirements would necessitate higher pay increases than would those of the Army.

If Army requirements impose a pay increase on the Navy which is greater than the Navy requires to maintain 70 percent of total enlistments in AFQT Categories I, II, and IIIa, then the overall AFQT profile of Navy enlistments will shift upward. Conversely, if pay increases determined from Army requirements are lower than those determined from Navy requirements (e.g. when required Navy accession are 135,000), then the AFQT profile of Navy accessions will shift downward. Figure 10 illustrates this point, indicating the percentage of total USN enlistments in AFQT Categories I, II, and IIIa for various Army accession requirements—assuming that Army requirements are the driving force.

There is of course a third possibility, not yet considered here. Pay increases may not be sufficient to meet either the needs of the Army or the Navy. Figures 11 and 12 illustrate the quality effect of various pay increases on USN FY '72, '73 accessions in a draft-free environment. These estimates are based on methodology which may be found in Ref. [7, Annex D].

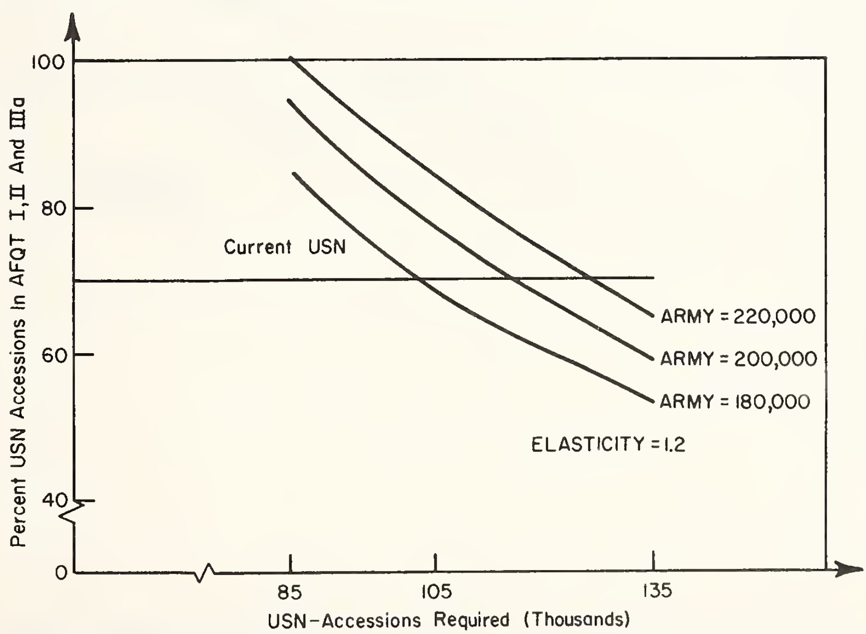


FIGURE 10. "Optimistic" percent USN true volunteers in AFQT I, II, and IIIA vs USN accession levels for various Army Accession level requirements (assuming that Army requirements determine pay increases for all services).

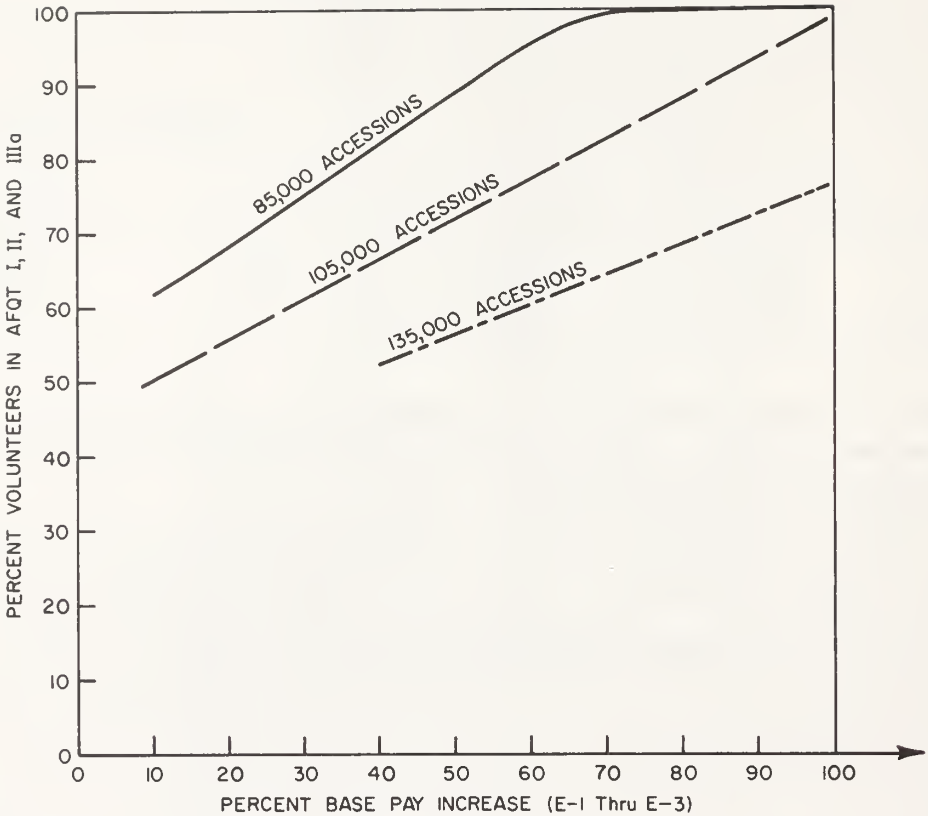


FIGURE 11. The effects of base pay increases on USN accessions in FY–'72, '73 (optimistic estimates).

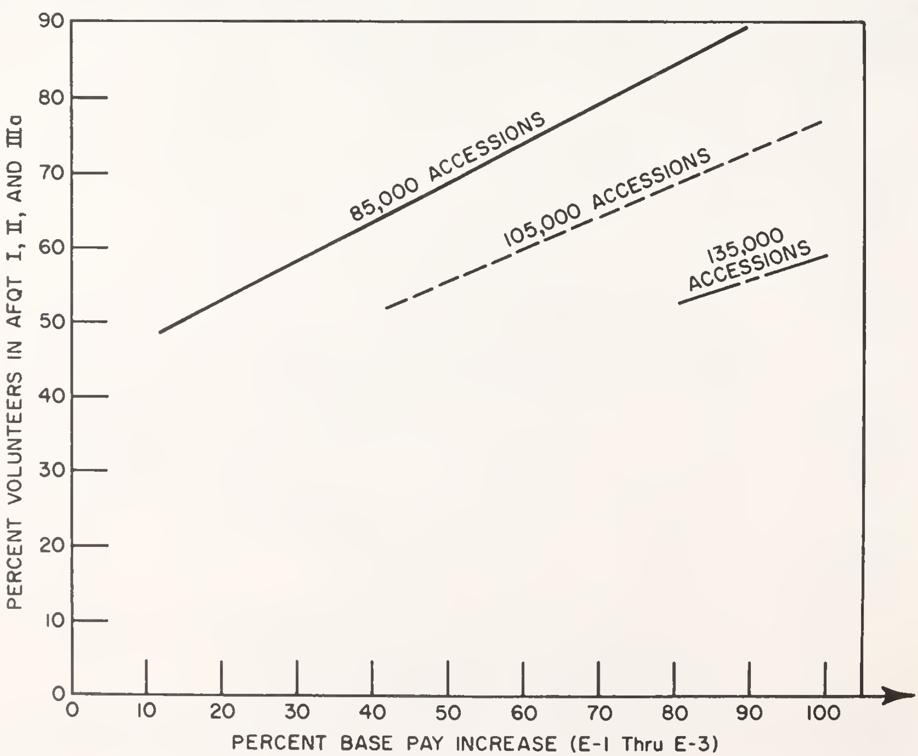


FIGURE 12. The effects of base pay increases on USN accessions in FY–'72, '73 (pessimistic estimates).

ACKNOWLEDGEMENT

The authors wish to express their appreciation to Mr. Robert Lehto, BUPERS (Pers A–12) and Mr. Burton Gray, Center for Naval Analyses (CNA) for their informal advice; to LCDR D. Y. Acosta, BUPERS (Pers N211) for her assistance in gathering statistical data; and to YN1 Jerry Leverett for his valuable research assistance.

REFERENCES

[1] Fechter, A. E. and S. A. Hoenack, "Econometric Models of Enlistments and Re-Enlistments in the U.S. Army," Institute for Defense Analyses (June 1968).

[2] Fechter, A. E. and S. H. Altman, "The Supply of Military Personnel in the Absence of a Draft," American Economic Review (May 1967), pp. 19–31.

[3] Gray, B. C., "The Supply of First Term Military Enlistees A Cross-Section Analysis," Center for Naval Analyses (July 1970).

[4] Karpinos, B. D., "The Mental Qualification of American Youths for Military Service and Its Relationship to Educational Attainment," Social Statistics Section, Proceedings of the American Statistical Assoc. (1966).

[5] Oi, W. Y., "The Cost and Implications of an All Volunteer Force," in Sol Tax (ed.), *The Draft: A Handbook of Facts and Alternatives* (University of Chicago Press, Chicago, 1967).

[6] O'Neill, D., "Forecasts of the Supply of Navy Enlisted First-Term Accessions in the Absence of a Draft: Fiscal Year 1970 to 1980," Center for Naval Analyses (June 1969).

[7] Rhode, A. S., J. J. Gelke, and F. X. Cook, "Impact of An All Volunteer Force Upon The Navy in the 1972–1973 Timeframe," Office of the Chief of Naval Operations (Dec. 1970), Annexes A–E.

[8] Rosenbaum, D. E., "Drawing Tonight Will Determine Who is Drafted," New York Times (1 Dec. 1969).

[9] Rosenbaum, D. E., "Lottery is Held to Set the Order of Draft in 1970," New York Times (2 Dec. 1969).

[10] Rosenbaum, D. E., "2nd Draft Lottery Selects Call-up Order for 1971," New York Times (2 July 1970).

THE MULTICOMMODITY NETWORK FLOW MODEL REVISED TO INCLUDE VEHICLE PER TIME PERIOD AND NODE CONSTRAINTS

Henry S. Weigel and John E. Cremeans

Research Analysis Corporation

ABSTRACT

The minimum-cost formulation of the problem of determining multicommodity flows over a capacitated network subject to resource constraints has been treated in previous papers. In those treatments only capacitated arcs were assumed and a uniform unit of measure like short tons was used for all commodities. This paper treats the effect of constraints on the nodes of the network, allows the commodities to be measured in their "natural" units and allows the network capacities to be expressed in vehicles per time period—in some cases giving a more accurate representation of the capacities of the network. This paper describes the solution procedure which uses the column generation technique; it also discusses computational experience.

INTRODUCTION

In a previous paper [1] the maximum flow multicommodity network problem, as formulated by Ford and Fulkerson [2], and the corresponding minimum cost problem, as stated by Tomlin [3], were extended to include resource constraints and resource substitution. The purpose of this paper is to describe other modifications to the multicommodity network flow model that have been found to be useful in studies of transportation networks.

These modifications will be described in two steps. First, the problem will be revised to permit the expression of commodity flows in terms of the natural unit of the commodity and the expression of arc capacities in terms of vehicles per time period. In the standard formulation, flows and requirements for commodity movement are converted into some common unit, such as short tons, for processing by the algorithm. It is desirable, both for convenience and accuracy, to represent all commodities in terms of their natural units (e.g., number of passengers, barrels of gasoline, tons of coal, etc.). It is also convenient to express arc capacities in terms of vehicles per time period since most traffic statistics are collected in this fashion.

Second, node flow constraints will be added to the problem formulation. In the analysis of transportation networks, there are frequently cases where the intersection of three or more arcs has itself a capacity. The capacity of a node is not easily expressed through the addition of dummy arcs. It will be shown that node constraints can be added to permit explicit statement of these capacities.

MINIMUM COST MULTICOMMODITY NETWORK FLOWS WITH RESOURCE CONSTRAINTS

For sake of continuity, the minimum cost problem with resource constraints is restated.

Consider the multimode, multicommodity network $G(N, \mathscr{A})$. N is the set of all nodes of the net-

work. $\mathscr{A}$ is the subset of all ordered pairs (x, y) of the elements of N that are arcs of the network. $\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_m$ is an enumeration of the arcs. Each arc has an associated cost (or distance) $d(x, y) \geq 0$ and each arc has an associated capacity $b(x, y) \geq 0$.

For each commodity $k (k=1, \ldots, q)$ there is a source, s_k, and a sink, t_k. The flow of the commodity, k, over arc (x, y) is denoted by $F^k(x, y)$, $k=1, \ldots, q$. The flows and the arc capacities must be stated in terms of some common unit such as short tons. These flows must satisfy the capacity constraints.

$$\sum_{k=1}^{q} F^k(x, y) \leq b(x, y), \qquad (x, y) \epsilon \mathscr{A}.$$

In keeping with previous notation, define the set $P^k = \{P_j^{(k)} | P_j^{(k)}$ is a chain connecting s_k and $t_k\}$. Now let P be the union of the sets $P^k (k=1, \ldots, q)$. Further, let $P_1^{(1)}, P_2^{(1)}, \ldots, P_j^{(k)}, \ldots, P_n^{(q)}$ be the enumeration of the chains $P_j^{(k)} \epsilon P$ such that the subscript j is sufficient to identify the chain, its origin-destination pair, and the commodity with which it is associated.

Thus the kth commodity set is defined by

$$J_k = \{j | P_j^{(k)} \text{ is a chain from } s_k \text{ to } t_k\}, \; k=1 \ldots, q.$$

The arc-chain incidence matrix is

$$A = [a_{ij}],$$

where

$$a_{ij} = \begin{cases} 1, & \text{if } \mathscr{A}_i \epsilon P_j^{(k)} \\ 0, & \text{otherwise} \end{cases}$$

for $i=1, \ldots m$; $j=1, \ldots, n$.

Define the resource matrix for commodity k by

$$R^k[r_{is}^k] (i=1, \ldots, m; \; s=1, \ldots, p),$$

where r_{is}^k is the quantity of resource s required to sustain a unit of flow of commodity k over arc i; $r_{is}^k \geq 0$. And let $\rho_s (s=1, \ldots, p)$ be the quantity of resource s available (e.g., in inventory) for assignment to the network.

The objective is to minimize the total cost of the resources used and the tolls on the arcs while meeting delivery requirements and satisfying the arc capacity and resource inventory constraints. Letting $x_j^{(k)} (j=1, \ldots, n)$ be the flow of commodity k in chain $P_j^{(k)} (j=1, \ldots, n$; k implicit), b_i the flow capacity of $\mathscr{A}_i$, in arc-chain terms, the minimum cost network flow linear program with resource constraints is:

Minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)} = z$$

subject to

(a) arc capacity constraints

$$\sum_{j=1}^{n} a_{ij} x_j^{(k)} \leq b_i, \; i=1, \ldots, m,$$

(b) resource constraints

$$\sum_{i=1}^{m} \sum_{k=1}^{q} \sum_{j\in J_k} a_{ij} r_{is}^{k} x_j^{(k)} \leq \rho_s, \; s=1, \; . \; . \; . \; , p,$$

and

(c) delivery requirements

$$\sum_{j\in J_k} x_j^{(k)} = \lambda_k, \; k=1, \; . \; . \; . \; , q,$$

where λ_k is the delivery requirement at t_k, $\lambda_k \geq 0$, $(k=1, \; . \; . \; . \; , q)$.

The cost coefficient c_j may be defined as:

$$c_j = \sum_{i=1}^{m} \tau_i a_{ij} + \sum_{s=1}^{p} \sum_{i=1}^{m} \Phi_s r_{is}^{k} \, a_{ij} \qquad (\text{for } j=1, \; . \; . \; . \; , n; \; k,$$

where $P_j^{(k)}$ connects s_k and t_k) and

where τ_i is the cost (or toll) for a unit flow over arc i, and Φ_s is the cost of using a unit of resource s.

THE REVISED PROGRAMMING PROBLEM

Use of Natural Units and Vehicles Per Time Period

The capacity $b(x, y)$ of arc $(x, y)\epsilon\mathscr{A}$ will now be expressed in vehicles-per-time-period. Some arcs will have vehicle types naturally associated with them. Highway, rail, waterway, and airway arcs will be naturally associated with trucks or busses, rail cars, barges, and aircraft, respectively. For other modes, such as pipeline and transfer, the "vehicle" must be contrived. One alternative is to use a dummy vehicle which contains exactly one unit of the commodity. All vehicles associated with a particular transportation mode are assumed to travel at a uniform rate of speed so that the arc capacities of a given mode are not dependent on the types of vehicles allowed to traverse the arcs of that mode. (This is also applicable to node constraints, which will be discussed later.)

It should be stressed at this point that all commodities will be counted in their natural unit. That is petroleum products will be expressed in barrels, passengers as numbers of passengers, and bulk cargo in tons if these are the units best suited to them. Throughout the remainder of this paper each reference to flow or movement of a commodity will be in the natural unit of that commodity.

To express the flow $F^k(x, y)$ in the natural unit of commodity k and the arc constraints in vehicle-per-time-period, rewrite the capacity constraints as follows:

$$\sum_{k=1}^{q} \frac{F^k(x, y)}{U^k(x, y)} \leq b(x, y), (x, y)\epsilon\mathscr{A};$$

where $U^k(x, y)$ is the quantity of the commodity k transported over arc (x, y) in a single vehicle, $U^k(x, y) > 0$; and $b(x, y)$ is now in vehicles per time period.

Define a matrix of vehicle per commodity unit coefficients as:

$$V = [v_i^k],$$

where

$$v_i^k = \frac{1}{U^k(x, y)} \quad \text{(the arc } (x, y) \text{ is denoted as the } i\text{th arc by the index } i).$$

In words, v_i^k is the reciprocal of the quantity of commodity k transported over arc i in a single vehicle.

With the above change, the minimum cost linear program can now be stated as follows:

Minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)} = z$$

subject to

(a) arc capacity constraints

$$\sum_{j=1}^{n} a_{ij} v_i^k x_j^{(k)} \leq b_i, \; i=1, \ldots, m,$$

(b) resource constraints

$$\sum_{i=1}^{m} \sum_{k=1}^{q} \sum_{j \epsilon J_k} a_{ij} r_{is}^k x_j^{(k)} \leq \rho_s, \; s=1, \ldots, p$$

and

(c) delivery requirements

$$\sum_{j \epsilon J_k} x_j^{(k)} = \lambda_k, \; k=1, \ldots, q.$$

Node Constraints

Node constraints allow a direct limitation on the throughput capacity of a node without the necessity of additional assumptions in depicting the node graphically as a small subnetwork. Examples may clarify this point.

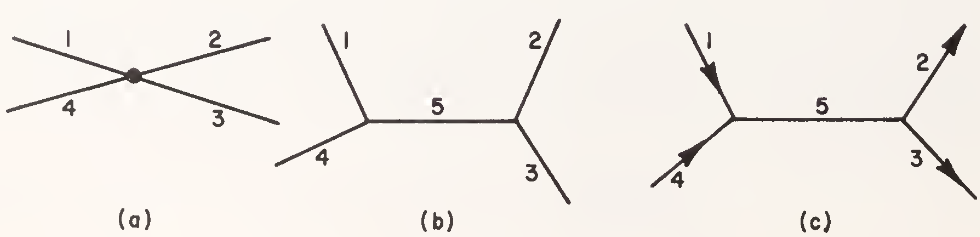


FIGURE 1.

In Figure 1, the node given in (a) is to be constrained by replacing it with the subnetwork given in (b). The capacity limitation on arc 5 is that of the node it has replaced. Note, however, that the path 1–4 in (a) contributes to the flow through the node, but in (b) it does not contribute to the flow on arc 5. In some cases, this problem may be solved by directing the arcs, i.e., by permitting flow in only one direction as in Figure 1–c. This method is not always satisfactory because in a multi-commodity problem

the optimum solution may require that one commodity flow over path 1–4 while another is required to flow over path 2–4.

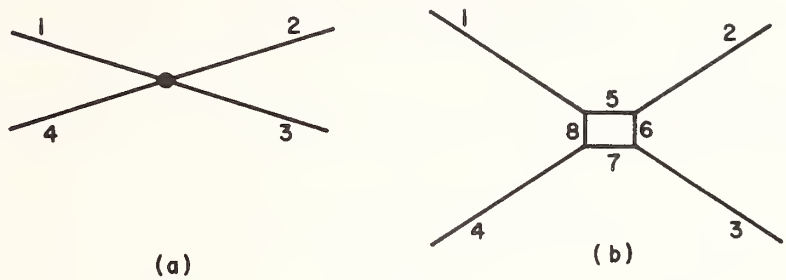


FIGURE 2.

A representation such as that shown in Figure 2 is sometimes attempted, but this too is unsatisfactory. What capacity should be assigned to arcs 5, 6, 7, and 8?

To get an accurate node throughput limitation, a constraint row is added in the problem formulation for each node for which a capacity limitation is desired. Let $\hat{N}$ be the subset of N of those nodes which are to be constrained. To consider the node throughput limitations, let $N_1, \ldots, N_t$ be an enumeration of the nodes $N_l \epsilon \hat{N}$. Let $\hat{b}_l$ be the capacity associated with the node N_l which is expressed in vehicles per time period.

Define a matrix of vehicle per commodity unit coefficients for node constraints similar to those for the arc constraints:

$$V = [\hat{v}_l^k],$$

where

$$\hat{v}_l^k = \frac{1}{\hat{U}^k(N_l)}$$

and $U^k(N_l)$ is the quantity of the commodity k transported through node N_l in a single vehicle, $U^k(N_l) > 0$. Let $\hat{b}_l$ be the throughput capacity of node N_l.

The node-chain incidence matrix is $F = [f_{lj}]$,

where

$$f_{lj} = \begin{cases} 1, & \text{if } N_l \epsilon P_j^{(k)} \\ 0, & \text{otherwise} \end{cases}$$

for $l = 1, \ldots, t$; $j = 1, \ldots, n$.

In arc-chain terms, the minimum cost network flow linear program with resource constraints, node constraints and putting the capacity constraints in vehicles per time is as follows:

Minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)} = z,$$

subject to

(a) arc capacity constraints

$$\sum_{j=1}^{n} a_{ij} v_i^k x_j^{(k)} \leq b_i, \; i = 1, \ldots, m,$$

(b) node capacity constraints

$$\sum_{j=1}^{n} f_{lj} \hat{v}_l^k x_j^{(k)} \leq \hat{b}_l, \; l = 1, \ldots, t,$$

(c) resource constraints

$$\sum_{i=1}^{m} \sum_{k=1}^{q} \sum_{j \epsilon J_k} a_{ij} r_{is}^k x_j^{(k)} \leq \rho_s, \; s = 1, \ldots, p,$$

and

(d) delivery requirements

$$\sum_{j \epsilon J_k} x_j^{(k)} = \lambda_k, \; k = 1, \ldots, q.$$

SOLUTION PROCEDURE USING THE COLUMN GENERATION TECHNIQUE

We have defined a straight forward linear programming problem different from that previously defined only in that node constraints have been added, the natural units are used for the flows and the delivery requirements, and vehicles per time period are used as capacities. It is now necessary to show that the column generation technique can be adapted to these changes. This technique is more fully described in Reference 1.

Define the matrices A^*, E, F^*, and G as follows:

$$A^* = [a_{ij}^*],$$

where

$$a_{ij}^* = a_{ij} v_i^k, \qquad k \text{ implied by } P_j^{(k)}.$$

$$E = [e_{sj}],$$

where

$$e_{sj} = \sum_{i=1}^{m} r_{is}^k a_{ij}, \qquad k \text{ implied by } P_j^{(k)}.$$

$$F^* = [f_{lj}^*],$$

where

$$f_{lj}^* = f_{lj} \hat{v}_l^k, \qquad k \text{ implied by } P_j^{(k)},$$

and

$$G = [g_{kj}]$$

where

$$g_{kj} = \begin{cases} 1, & \text{if } j \epsilon J_k \\ 0, & \text{otherwise} \end{cases}$$

$\hat{A}$ is a matrix $(m+p+t+q \times n)$ formed of the submatrices A^*, E, F^*, and G as follows:

$$\hat{A} = \begin{bmatrix} A^* \\ \hline E \\ \hline F^* \\ \hline G \end{bmatrix}$$

The typical column of $\hat{A}$ is

$$\hat{A}_j = \text{col.}\ (a^*_{1j}, \ldots, a^*_{mj}, e_{1j}, \ldots, e_{pj}, f^*_{1j}, \ldots, f^*_{tj}, g_{1j}, \ldots, g_{qj}).$$

$\hat{A}$ will be quite large due to the fact that there is typically an enormous number of chains connecting s_k and t_k $(k=1, \ldots, q)$. The shortest chain algorithm can be used to develop the $\hat{A}_j$ that will satisfy the simplex rule. Further if the shortest chain algorithm can find no chain satisfying the requirement, an optimum has been reached.

This formulation can be solved by adopting the standard two-phased procedure. Phase I minimizes to zero the value of

$$\sum_{j=n+m+t+p+1}^{j=n+m+t+p+q} x_j^{(k)}$$

to obtain an initial basic feasible solution. This effectively assigns a cost of 1 to the artificial variables and a cost of zero to the other variables in Phase I. Phase II begins with the basic feasible solution determined in Phase I and proceeds to minimize

$$\sum_{j=1}^{n} c_j x_j^{(k)} = z.$$

In Phase I, $I_{m+t+p+q}$ may be used as the initial basis and the simplex rule is to enter a chain in the basis if, and only if,

$$c_j - c_B B^{-1} \hat{A}_j < 0$$

where

$$c_B B^{-1} = (\alpha_1, \ldots, \alpha_m, \pi_1, \ldots, \pi_p, \beta_1, \ldots, \beta_t, \sigma_1, \ldots, \sigma_q)$$

so that the simplex multipliers α_i are associated with the arcs, the π_s are associated with the resources,

the β_l are associated with the nodes, and the σ_k are associated with the artificial variables. Thus the vector $\hat{A}_j$ is entered if

$$-\sum_{i=1}^{m} a_{ij}\left[\alpha_i v_i^k + \sum_{s=1}^{p} \pi_s r_{is}^k\right] - \sum_{l=1}^{t} f_{lj}\hat{v}_i^k \beta_l < \sigma_k.$$

The contribution of each arc to $c_j - c_B B^{-1}\hat{A}_j$ in Phase I is

$$d_i^k = -\left[\alpha_i v_i^k + \sum_{s=1}^{p} \pi_s r_{is}^k\right]; \text{ noting that } c_j = 0 \text{ for structural vectors.}$$

The contribution of each node to $c_j - c_B B^{-1}\hat{A}_j$ must be distributed over the arcs to facilitate the use of the shortest chain algorithm. Disallowing the origin and destination nodes from being constrained, it is clear that if a constrained note is contained in a route, the route also contains precisely two of the arcs which emanate from or terminate at this node. This suggests a method for the distribution of the node contributions. Namely, that half of the contribution be added to each arc which has this node as an end point.

To do this, define

$$\delta_{li} = \begin{cases} 1, & \text{if node } N_l \text{ is an end point of arc } \mathscr{A}_i \\ 0, & \text{otherwise.} \end{cases}$$

The effective contribution of.each arc to $c_j - c_B B^{-1}\hat{A}_j$ becomes

$$\hat{d}_i^k = -\left[\alpha_i v_i^k + \sum_{s=1}^{p} \pi_s r_{is}^k + \tfrac{1}{2}\sum_{l=1}^{t} \delta_{li}\hat{v}_i^k \beta_l\right].$$

The shortest chain algorithm may be used to find

$$\min_j\left[\sum_{\mathscr{A}_i \epsilon P_j^{(k)}} \hat{d}_i^k\right] = \min_j\left[\sum_{\mathscr{A}_i \epsilon P_j^{(k)}} \left(-\alpha_i v_i^k - \sum_{s=1}^{p} \pi_s r_{is}^k - \tfrac{1}{2}\sum_{l=1}^{t} \delta_{li}\hat{v}_i^k \beta_l\right)\right]$$

over all k.

Where $\mathscr{A}_i$ is the ith arc and $P_j^{(k)}$ is the jth chain from s_k to t_k. The minimum over all commodities is selected as the candidate to enter the basis. The column vector to leave the basis may be determined in the standard simplex fashion. Should any α_i, π_s, or β_l be positive the corresponding slack variable must be entered into the basis.

In Phase II

$$c_j - c_B B^{-1}\hat{A}_j = \sum_{i=1}^{m} \tau_i a_{ij} + \sum_{s=1}^{p} \Phi_s e_{sj} - \sum_{i=1}^{m} v_i^k \alpha_i a_{ij}$$

$$- \sum_{s=1}^{p} \pi_s e_{sj} - \sum_{l=1}^{t} f_{lj}\hat{v}_i^k \beta_l - \sum_{w=1}^{q} \sigma_w g_{wj} = \sum_{i=1}^{m} a_{ij}(\tau_i - v_i^k \alpha_i)$$

$$+\sum_{s=1}^{p}\sum_{i=1}^{m} a_{ij}r_{is}^{k}(\Phi_s-\pi_s)-\sum_{l=1}^{t} f_{lj}\hat{v}_l^k\beta_l-\sum_{w=1}^{q}\sigma_w g_{wj},$$

where

$$e_{sj}=\sum_{i=1}^{m} r_{is}^{k}\, a_{ij}$$

$$g_{kj}=\begin{cases}1, & \text{if } j\epsilon J_k\\ 0, & \text{otherwise,}\end{cases}$$

and τ_i is the cost (or toll) per unit flow over arc i, as before, but the unit flow is now measured in the commodities' natural units rather than short tons.

Thus

$$d_i^k=\tau_i-v_i^k\alpha_i+\sum_{s=1}^{p} r_{is}^{k}(\Phi_s-\pi_s)$$

and

$$\hat{d}_i^k=\tau_i-v_i^k\alpha_i+\sum_{s=1}^{p} r_{is}^{k}(\Phi_s-\pi_s)-\tfrac{1}{2}\sum_{l=1}^{t}\delta_{li}\hat{v}_l^k\beta_l$$

may be assigned to arc i. The shortest, i.e., the chain with the least

$$\sum_{i=1}^{m}\hat{d}_i^k-\sigma_k<0$$

for $k=1, \ldots, q$, may then be entered in the basis.

Phase II is terminated and the value of z is minimized when, for the minimum j,

$$\sum_{\mathcal{A}_i\epsilon P_j^{(k)}}\hat{d}_i^k \geqslant \sigma_k.$$

MODIFICATION OF THE SUBSTITUTION PROCEDURE

The substitution procedure is affected by the change in that the use of a different group of resources may require that a different coefficient of vehicles per commodity unit be used. Under the procedure described above the matrix V is a two dimensional matrix by arc and by commodity. When substitution is permitted this matrix must be given the additional dimension of substitution group.

Redefine the matrices V and $\hat{V}$, the matrices of vehicles per commodity unit, as:

$$V=[v_{ih}^k] \quad \text{and} \quad \hat{V}=[\hat{v}_{lh}^k], \quad \text{respectively,}$$

where v_{ih}^k is the reciprocal of the quantity of commodity k transported over arc i in a single vehicle determined by substitution group h and $\hat{v}_{lh}^k$ is the reciprocal of the quantity of commodity k transported through node l in a single vehicle determined by substitution group h. The coefficients v_{ih}^k and $\hat{v}_{lh}^k$ are

therefore selected in each iteration based on the minimum "cost" (in terms of simplex multipliers) alternative resource vector to be used.

Summary of the procedure

To summarize, the proposed procedure is:

(1) Calculate $c_B B^{-1} = (\alpha_1, \ldots, \alpha_m, \pi_1, \ldots, \pi_p, \beta_1, \ldots, \beta_t, \sigma_1, \ldots, \sigma_q)$.

(2) In the previous formulation, for each arc-commodity pair a single combination of resources is required and represented by the vector.

$$R_i^k = (r_{i1}^k, r_{i2}^k, \ldots, r_{ip}^k), \qquad \text{of the matrix } R^k.$$

Define a new resource matrix:

$$T = [t_{ik}]\,(i = 1, \ldots, m;\ k = 1, \ldots, q)$$

where

$$t_{ik} = \{\hat{R}_i^k \mid \hat{R}_i^k \text{ is any feasible resource vector for arc } i, \text{ commodity } k\}\ \hat{R}_i^k = (\hat{r}_{i1}^k, \ldots, \hat{r}_{ip}^k).$$

In words, each element of T is the set of alternative resource vectors for a movement of one unit of commodity k over arc i. For each arc-commodity pair find the least-cost applicable resource vector, "cost" meaning cost in terms of the simplex multipliers and resource prices and call it the vector h.

$$\bar{R}_{ih}^k = \left\{\hat{R}_i^k \,\middle|\, \sum_{s=1}^{p} \hat{r}_{is}^k(\Phi_s - \pi_s) \text{ is minimum for } \hat{R}_i^k \epsilon t_{ik}\right\},$$

where

$$\bar{R}_{ih}^k = (\bar{r}_{i1}^k, \ldots, \bar{r}_{ip}^k).$$

(3) For commodity $k = 1, \ldots$ q calculate

$$\hat{d}_i^k = \tau_i - \alpha_i v_{ih}^k + \sum_{s=1}^{p} \bar{r}_{is}^k(\Phi_s - \pi_s) - \tfrac{1}{2}\sum_{l=1}^{t} \delta_{li}\hat{v}_{ih}^k\beta_l \qquad \text{for } i = 1, \ldots, m$$

and assign $\hat{d}_i^k$ to arc i as a pseudo cost.

(4) Using the shortest chain algorithm, find the chain with the least

$$\hat{d}_j^k = \sum_{i=1}^{m} a_{ij}\left[\tau_i - \alpha_i v_{ih}^k + \sum_{s=1}^{p} \bar{r}_{is}^k(\Phi_s - \pi_s) - \tfrac{1}{2}\sum_{l=1}^{t} \delta_{li}\hat{v}_{ih}^k\beta_l\right] \qquad \text{for } k = 1, \ldots, q.$$

(5) Find $\min\limits_k\,[\hat{d}_j^k - \sigma_k]$.

(6) If the minimum $[\hat{d}_j^k - \sigma_k] < 0$, the vector $\hat{A}_j$ is entered in the basis. If minimum $[\hat{d}_j^k - \sigma_k] \geqslant 0$ there is no chain that may improve the value of the objective function, and the procedure is terminated.

CONCLUSIONS

It seems to be the consensus of the literature that the basic Ford-Fulkerson multicommodity algorithm attains its speed based, in part, upon the fact that the conventional method results in a zero-one matrix. It should be emphasized, however, that the literature contains no direct test of this idea. The assumption of speed seems to be based largely on the simplicity of the zero-one matrix and the fact that the decomposition algorithm (that is directly related to the Ford-Fulkerson suggested method) typically does require large amounts of computer time.

The Control Data 6400 computer program for the algorithm described in [1] was revised to include the changes discussed in this paper. It was of interest to make a comparison of the performances of the unrevised and the revised algorithms. Two problems were used as tests. One, a small problem with 52 arcs, 5 resources, 2 origin-destination pairs, and no node constraints, a total of 59 rows. The computer time statistics on the CDC 6400 of two variations of the problem formulation are given in Table 1 below. One formulation has the arc capacities, the movement requirements and the route flows in terms of tons per day. The other has the arc capacities in vehicles per day, the movement requirement and the route flows in the commodities' natural units.

TABLE 1. *Computer Time Statistics for Problem 1*

Constraint type	Central processor time (sec)	Number of iterations	Time per iteration (sec)
Vehicles per day	13.3	19	0.70
Tons per day	11.8	23	0.51

In this simple problem one arc constraint type was directly converted from the other to make the problems as much the same as possible. As a result, the objective function values were the same and the flows over respective arcs were the same. Every effort was made to keep the problems similar so that the timing comparisons would be meaningful. The central processor time for the run with the arc constraints in vehicles per day was 12.7 percent greater than that for the run with the arc constraints in tons per day.

The second problem which was used as a test did not have the exact conversion from the tons formulation to the vehicles formulation as the first one had.* In the vehicles formulation two arcs of the network were combined into one with the capacity being adjusted accordingly and several resources with identical productivities were combined and the resource inventories were adjusted accordingly. The tons formulation of the problem has 290 arcs, 42 resources, 9 origin-destination pairs (2 of them had zero requirements), a total of 341 rows. The vehicles formulation has the same statistics except for one less arc and five less resources. Table 2 gives the run time comparisons on the CDC 6400. As a result of the inexact conversion of the tons formulation to the vehicles formulation, the objective function for the vehicles formulation was slightly smaller, as expected, and the solution was slightly different from the tons formulation. The central processor time for the run with the arc constraints in vehicles per day was 55 percent greater than that for the run with the arc constraints in tons per day.

TABLE 2. *Computer Time Statistics for Problem 2*

Constraint type	Central processor time (sec)	Number of iterations	Time per iteration (sec)
Vehicles per day	1860	265	7.0
Tons per day	1200	156	7.7

*This problem was solved for purposes other than those of testing the algorithm and the time to solve.

The revised algorithm will increase the accuracy in the handling of capacity limitations if it is assumed that a vehicle limitation is a more accurate expression of capacity than a limitation expressed in short tons. Since most capacities expressed in short tons are simple conversions from vehicle per time period figures using the average vehicle load, it seems likely that this is true.

The inclusion of the node constraints allows for a direct limitation on the throughput capacity of a node, rather than the more complex way of depicting a node graphically as a small subnetwork. "Networking" a node is possible only in specialized cases, or additional assumptions like directing the arcs need to be made. These additional assumptions tend to unduly restrict the network.

The revision also permits much more flexible modeling of specialized processes, such as transfer, depot, and other intra-nodal functions. The capacities of arcs representing such processes do vary depending on the commodity involved. The revised procedure will permit a more realistic handling of these problems.

REFERENCES

Cited References

[1] Cremeans, J. E., R. A. Smith, and G. R. Tyndall, "Optimal Multi-Commodity Network Flows with Resource Allocation," Nav. Res. Log. Quart., *17*, 269–279 (1970).

[2] Ford, L. R., Jr., and D. R. Fulkerson, "A Suggested Computation for Maximal Multi-Commodity Network Flows," Mgt. Sci. (Oct. 1958).

[3] Tomlin, J. A., "Minimum-Cost Multi-Commodity Network Flows," Operations Research (Dec. 1966).

Additional References

Boyer, Donald D., "A Modified Simplex Algorithm for Solving the Multi-Commodity Maximum Flow Problem," TM14930, The George Washington University Logistics Research Project, Washington, D.C. (Mar. 1968).

Busacker, R. G., et al., "Three General Network Flow Problems and Their Solutions," RAC–SP–183, Research Analysis Corporation, (Nov. 1962).

Busacker, R. G. and T. L. Saaty, *Finite Graphs and Networks* (McGraw-Hill Book Company, New York, N.Y., 1965).

Cremeans, J. E., "The Bridges of Konigsburg," ASTME Vectors (May–June 1969), p. 4.

Charnes, A., "Optimality and Degeneracy in Linear Programming," Econometrica, Vol. 20 (Apr. 1952).

Dreyfus, S. F., "An Appraisal of Some Shortest-Path Algorithms," The Rand Corp., RM–5433–1–PR (Sept. 1968).

Fitzpatrick, G. R., et al., "Programming the Procurement of Airlift and Sealift Forces: A Linear Pro-

gramming Model for Analysis of the Least Cost Mix of Strategic Deployment Systems," Nav. Res. Log. Quart. *14*, 241–255 (1967).

Ford, L. R., Jr. and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, N.J. (1962).

Gass, S. I., *Linear Programming—Methods and Applications* (McGraw-Hill Book Co., Inc., New York, 1958).

Hadley, G., *Linear Programming* (Addison-Wesley Publishing Company, Inc., Reading, Mass., 1962).

Koopmans, T. C., ed., *Activity Analysis of Production and Allocation* (John Wiley and Son, Inc., New York, N.Y., 1951).

Orchard-Hays, W., *Advanced Linear-Programming Computing Techniques* (McGraw-Hill Book Co., Inc., New York, 1968).

Rao, M. R. and S. Zionts, "Allocation of Transportation Units to Alternative Trips—A Column Generation Scheme with Out-of-Kilter Subproblems," Operations Research (Jan.–Feb. 1968).

Sakarovitch, M., "The Multi-Commodity Maximum Flow Problem," Operations Research Center, University of California, Berkeley (Dec. 1966).

AN EXTENSION OF THE (SZWARC) TRUCK ASSIGNMENT PROBLEM

Mandell Bellmore

Department of Operations Research
The Johns Hopkins University

Jon C. Liebman*

Department of Geography and Environmental Engineering
The Johns Hopkins University

and

David H. Marks

Department of Civil Engineering
Massachusetts Institute of Technology

ABSTRACT

In this journal in 1967, Szwarc presented an algorithm for the optimal routing of a common vehicle fleet between m sources and n sinks with p different types of commodities. The main premise of the formulation is that a truck may carry only one commodity at a time and must deliver the entire load to one demand area. This eliminates the problem of routing vehicles between sources or between sinks and limits the problem to the routing of loaded trucks between sources and sinks and empty trucks making the return trip. Szwarc considered only the transportation aspect of the problem (i.e., no intermediate points) and presented a very efficient algorithm for solution of the case he d·scribed. If the total supply is greater than the total demand, Szwarc shows that the problem is equivalent to a $(mp+n)$ by $(np+m)$ Hitchcock transportation problem. Digital computer codes for this algorithm require rapid access storage for a matrix of size $(mp+n)$ by $(np+m)$; therefore, computer storage required grows proportionally to p^2. This paper offers an extension of his work to a more general form: a transshipment network with capacity constraints on all arcs and facilities. The problem is shown to be solvable directly by Fulkerson's out-of-kilter algorithm. Digital computer codes for this formulation require rapid access storage proportional to p instead of p^2. Computational results indicate that, in addition to handling the extensions, the out-of-kilter algorithm is more efficient in the solution of the original problem when there is a mod·rate number of commodities and a computer of limited storage capacity.

INTRODUCTION

The problem of optimally routing vehicles which carry one or more products between sources of supply and areas of demand is of considerable interest and difficulty. Depending on the assumptions made about the capacity of the vehicles in relation to the size of demands, the uniformity and types of demands, the way demand is spread over the network to be served and the presence of additional constraints, very different problem formulations may be developed, with resulting differences in solution technique. Szwarc [6] formulated the "Truck Assignment Problem" which is actually a multi-commodity transportation problem with some special conditions. In this formulation, there exist m production sites each capable of producing p commodities, and n demand locations where there is a demand for each of the commodities. The demand for a commodity is such that the entire capacity of a

*Present address: Department of Civil Engineering, University of Illinois at Urbana.

vehicle is allocated to only one commodity and is destined for one demand location only. Thus the problem of routing a vehicle between supply locations or demand locations is avoided and a vehicle's assignment will consist of loading a commodity at a production source, delivering its contents to one demand location and returning empty to a production source for reassignment. The objective is to find the vehicle assignments that will minimize the total distance traveled while meeting the supply-demand constraints.

This problem, because of the requirements that a vehicle traveling from source to sink may carry only one commodity and that the return from the sink to the source is done empty, is a far less complicated problem than the general multicommodity flow problems considered by Appelgren [1], Levin [5], and Bellmore, Bennington, and Lubore [2]. Each arc in the general multicommodity flow problem may carry many different commodities at the same time. But with the special formulation, each arc from a source of a commodity to a sink of a commodity may carry only that commodity. Thus the problem may be decomposed into a series of single commodity problems in which loaded vehicles are shipped from sources to sinks, and empty vehicles from sinks back to sources. Arcs representing flow between sources and sinks of different commodities are not allowed. However, arcs from the sinks back to the sources represent empty trucks and any source may be reached from any sink.

Szwarc gave an algorithm based on solving one or more transportation problems in an enlarged network in which a source of several commodities becomes a distinct source for each of the commodities. Sinks are treated in the same manner. Szwarc distinguished between two different cases, with different algorithms for each. In the case where demand for each commodity is exactly equal to supply, the problem is decomposed into several transportation problems (one for each commodity and one for the empty returning trucks). If there are m sources and n sinks, each of these problems is of size $m \times n$, with possible reductions in the size of some problems if some sources or sinks do not provide or demand some commodities. When supply is not equal to demand, a p-commodity problem is solved as a single large transportation problem of size $(mp + n)$ by $(np + m)$.

MATHEMATICAL FORMULATION AND EXTENSION

A mathematical statement of Szwarc's problem may be constructed with the following definitions:

x_{ijk} = the number of vehicle loads of commodity k to be sent from production source i to demand site j,

x_{ji}^* = The number of empty vehicles sent from demand site j back to production source i,

c_{ijk} = cost of shipping one vehicle load of commodity k from i to j,

c_{ji}^* = cost of returning one empty vehicle from demand site j to production source i,

D_{jk} = demand for commodity k at demand site j in vehicle loads,

S_{ik} = supply of commodity k at production source i in vehicle loads,

m = number of production sources,

n = number of demand sites,

p = number of commodities.

Then the general mathematical formulation for Szwarc's delivery problem is:

Minimize:

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{p} c_{ijk}x_{ijk} + \sum_{j=1}^{n}\sum_{i=1}^{m} c_{ji}^{*}x_{ji}^{*}, \tag{1}$$

subject to

(2) $$\sum_{j=1}^{n} x_{ijk} \leqslant S_{ik} \qquad \{i=1, 2, \ldots, m;\ k=1, 2, \ldots, p\},$$

(3) $$\sum_{i=1}^{m} x_{ijk} \geqslant D_{jk} \qquad \{j=1, 2, \ldots, n;\ k=1, 2, \ldots, p\},$$

and

(4) $$\sum_{k=1}^{p} \sum_{i=1}^{m} x_{ijk} = \sum_{i=1}^{m} x_{ji}^{*} \qquad \{j=1, 2, \ldots, n\},$$

and

(5) $$\sum_{k=1}^{p} \sum_{j=1}^{n} x_{ijk} = \sum_{j=1}^{n} x_{ji}^{*} \qquad \{i=1, 2, \ldots, m\},$$

where

(6) x_{ijk}, x_{ji}^{*} are nonnegative integers.

The objective function (1) is to minimize the cost of loaded trips from source to sink and of empty return trips. Inequality (2) requires that the amount of commodity k shipped from source i to all demand areas must not exceed the supply of commodity k at that source. Similarly, inequality (3) specifies that the total amount of commodity k shipped to j from all sources must be greater than or equal to the demand for commodity k at that site. Equations (4) and (5) are flow equations requiring respectively that the number of empty vehicles sent out of a demand site must equal the number of loaded vehicles arriving there, and that the number of loaded vehicles sent out of a source must equal the number of empty vehicles arriving there. Inherent in this formulation is that vehicles must return to a source after leaving a demand site. Therefore, the solution is a circulation, and the whole pattern may be repeated at specified time intervals. Specifically, if n_i vehicles start at source i at the beginning of the problem, then after all shipments are completed, n_i vehicles will be at source i.

Examination of the above formulation and the graph structure shown in Figure 1 (for a case with $m=n=p=2$) indicates that the problem may be solved directly by the out-of-kilter algorithm [4]. The computational difference between the two methods is shown by the fact that the Szwarc method requires the solution of three 2x2 transportation problems when demand equals supply, and one 6x6 transportation problem when demand does not equal supply. For a solution by the out-of-kilter algorithm a graph of $mn(p+1)+p(m+n)$ arcs and $(m+n)(p+1)$ nodes (or, for the example case, 20 arcs and 12 nodes) must be solved.

Once the problem has been placed in the form of an out-of-kilter algorithm, a number of useful extensions become apparent. Two of these are perhaps most significant:

1. It is possible to consider any sort of routing network with intermediate facilities. In this case, it may easily be shown that adding intermediate node equations to the model formulation still leaves a graph solvable by the out-of-kilter algorithm, but not by Szwarc's algorithm. Since many vehicle routing systems involve the consideration of intermediate points, such an extension is of considerable practical value.

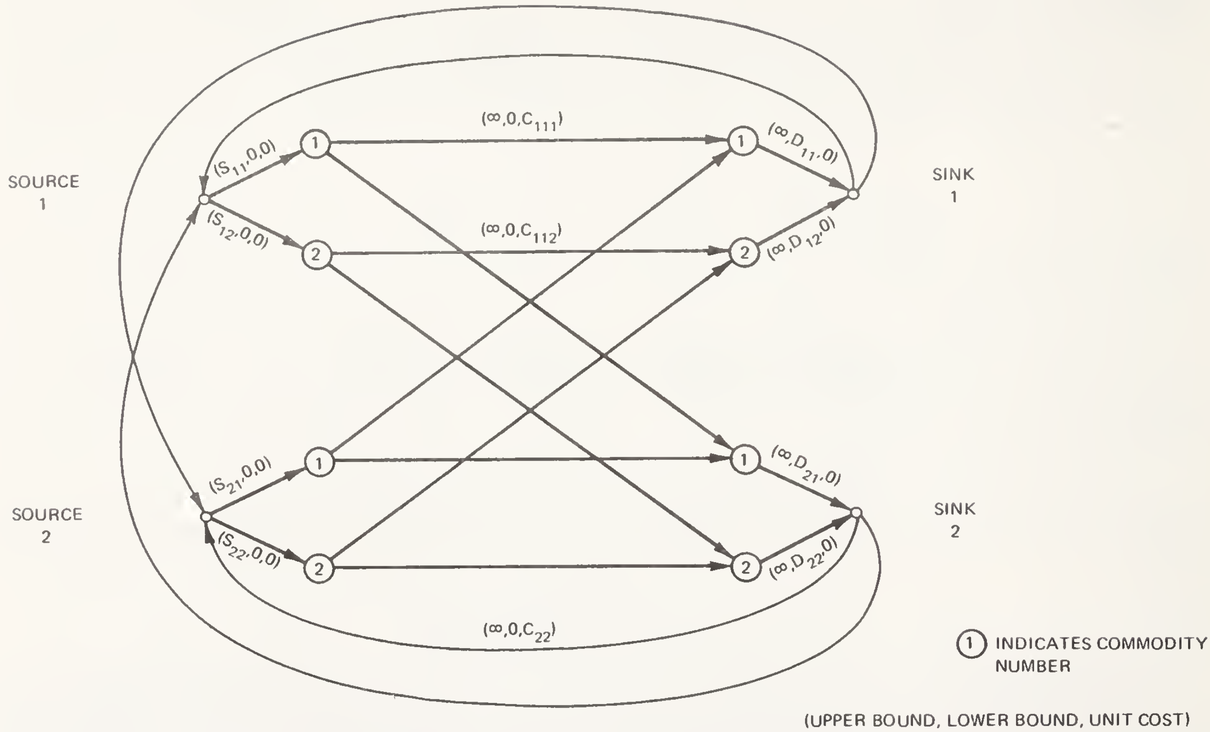


FIGURE 1. Graph representation of Szwarc truck assignment problem with $m=2$, $n=2$, and $p=2$

2. Many additional constraints may be added that will make the problem more realistic. Of particular interest are upper bounds on flows along particular routes of the form

$$x_{ijk} \leq Q_{ijk}, \tag{7}$$

where Q_{ijk} is the upper bound on flow through arc x_{ijk}. A finite upper bound might reflect a capacity constraint on the usage of a route. The out-of-kilter formulation will also permit imposition of lower-bound constraints on flow through the arcs, although these are unlikely to be needed in most problems.

We call this form of the truck assignment problem, with a set of f intermediate nodes that are neither sources nor sinks of commodities, the extended truck assignment problem:
Minimize

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} b_{ijk} x_{ijk} + \sum_{i=1}^{m} \sum_{a=1}^{f} \sum_{k=1}^{p} b^{*}_{iak} x^{*}_{iak} + \sum_{a=1}^{f} \sum_{j=1}^{n} \sum_{k=1}^{p} b^{**}_{ajk} x^{**}_{ajk} + \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ji} x_{ji}, \tag{8}$$

subject to

$$\sum_{j=1}^{n} x_{ijk} + \sum_{a=1}^{f} x^{**}_{iak} \leq S_{ik} \qquad \{i=1,2,\ldots,m;\, k=1,2,\ldots,p\}, \tag{9}$$

$$\sum_{j=1}^{n} x^{*}_{ajk} \leq V_{ak} \qquad \{a=1,2,\ldots,f;\, k=1,2,\ldots,p\}, \tag{10}$$

(11) $$\sum_{a=1}^{f} x^*_{ajk} + \sum_{i=1}^{m} x_{ijk} \geq D_{jk} \qquad \{j=1,2,\ldots,m;\ k=1,2,\ldots,p\},$$

(12) $$\sum_{i=1}^{m} x^{**}_{iak} = \sum_{j=1}^{n} x^*_{ajk} \qquad \{a=1,2,\ldots,f;\ k=1,2,\ldots,p\},$$

(13) $$\sum_{k=1}^{p}\sum_{i=1}^{m} x_{ijk} + \sum_{k=1}^{p}\sum_{a=1}^{f} x^*_{ajk} = \sum_{i=1}^{m} x_{ji} \qquad \{j=1,2,\ldots,n\},$$

(14) $$\sum_{k=1}^{p}\sum_{j=1}^{n} x_{ijk} + \sum_{k=1}^{p}\sum_{a=1}^{f} x^{**}_{iak} = \sum_{j=1}^{n} x_{ji} \qquad \{i=1,2,\ldots,m\},$$

(15) $$x_{ijk} \leq Q_{ijk} \qquad \{i=1,2,\ldots,n;\ j=1,2,\ldots,n;\ k=1,2,\ldots,p\},$$

(16) $$x^{**}_{iak} \leq Q^{**}_{iak} \qquad \{i=1,2,\ldots,m;\ a=1,2,\ldots,f;\ k=1,2,\ldots,p\},$$

(17) $$x^*_{ajk} \leq Q^*_{ajk} \qquad \{a=1,2,\ldots,f;\ j=1,2,\ldots,n;\ k=1,2,\ldots,p\},$$

(18) $$x_{ji} \leq Q_{ji} \qquad \{j=1,2,\ldots,n;\ i=1,2,\ldots,m\},$$

(19) $$x_{ijk}, \quad x^*_{ajk}, \quad x^{**}_{iak}, \quad x_{ji} \text{ are nonnegative integers,}$$

where

a = index relating to intermediate nodes.
i = index relating to supply points,
j = index relating to demand points,
k = index relating to commodities,
x_{ijk} = number of truck loads of commodity k sent directly from source i to sink j,
x^{**}_{iak} = number of truck loads of commodity k sent from source i to intermediate point a,
x^*_{ajk} = number of truck loads of commodity k sent from intermediate point a to sink j,
x_{ji} = number of empty trucks returned from sink j directly to source i,
b_{ijk} = unit cost of supplying demand for commodity k at sink j directly from source i, including costs of loading and unloading,
b^*_{iak} = unit cost of shipping to a as an intermediate point for commodity k from source i, including loading cost,
b^{**}_{ajk} = unit cost of using a as an intermediate point for supplying demand for commodity k at sink j including costs of transshipment and unloading,
c_{ji} = cost of shipping an empty truck from sink j to source i,
S_{ik} = supply in truck loads of commodity k at source i,
D_{jk} = demand in truck loads for commodity k at sink j,
Q_{ijk} = upper bound on flow of commodity k from i to j,
Q^{**}_{iak} = upper bound on flow of commodity k from i to a,
Q^*_{ajk} = upper bound on flow of commodity k from a to j,

Q_{ji} = upper bound on flow of vehicles from j to i,

V_{ak} = upper bound on transshipment of commodity k at intermediate point a.

Inequalities (9), (10), and (11) express the constraints, respectively, on the amount of each commodity shipped from i, transshipped through a, and received by j. Equations (12), (13), and (14) express the conservation of flow constraints at each point. Inequalities (15), (16), (17), and (18) restrict the flow between points to be less than a bound. The integer requirement (19) is satisfied by the solution of the problem as an out-of-kilter graph.* The construction of the required graph is completely described as follows:

1. For each source i, a set of $p+1$ nodes is created and the nodes are named $y_i, y_{i1}, \ldots, y_{ip}$, where y_{ik} represents the k'th commodity produced at the source i, and y_i represents the i'th source.
2. Directed arcs are drawn from y_i to each of the y_{ik}. The lower and upper bounds on capacity of the arcs are zero and S_{ik}. The unit cost is zero.
3. For each intermediate point a, construct $2p$ nodes $a^1, a^2, \ldots, a^p$ and $A^1, A^2, \ldots, A^p$. Construct p arcs (a^k, A^k), $k=1, 2, \ldots, p$, whose upper and lower bounds are V_{ak} and zero. The unit cost is zero.
4. For each sink j, a set of $p+1$ nodes is created and the nodes are named $q_j, q_{j1}, \ldots, q_{jp}$, where q_{jk} represents the k'th commodity received at sink j, and q_j represents the j'th demand. Directed arcs are drawn from each node q_{jk} to q_j. The lower and upper bounds on each arc are D_{jk} and ∞. The unit cost is zero.
5. Directed arcs are drawn from each node y_{ik} to each intermediate node a^k, with upper and lower bounds Q^*_{iak} and zero, and unit cost b^*_{iak}.
6. Directed arcs are drawn from each node y_{ik} to each node q_{jk}, with upper and lower bounds Q_{ijk} and zero, and unit cost b_{ijk}.
7. Directed arcs are drawn from each intermediate node A^k to each node q_{jk}, with upper and lower bounds Q^{**}_{ajk} and zero, and unit cost b^{**}_{ajk}.
8. Directed arcs are drawn from each node q_j to each node y_i with upper and lower bounds Q_{ji} and zero, and unit cost C_{ji}.

The out-of-kilter solution must be decoded to obtain the optimal routings. This decoding is accomplished as follows:

1. No feasible solution for the graph means no feasible solution for the problem.
2. The flow from node y_i to node y_{ik} represents the number of truck loads of commodity k supplied at source i.
3. The flow from node y_{ik} to a^k represents the number of truck loads of commodity k shipped from source i to intermediate point a.
4. The flow from node y_{ik} to q_{jk} represents the number of truck loads of commodity k shipped from source i to sink j.
5. The flow from A^k to q_{jk} represents the number of truck loads of commodity k shipped from intermediate point a to sink j.
6. The flow from node q_j to node y_i represents the number of empty trucks returned from sink j to source i.
7. The cost of the solution to the graph is the cost of the solution to the delivery problem and is optimal.

* It is assumed that all coefficients in the problem formulation are nonnegative integers.

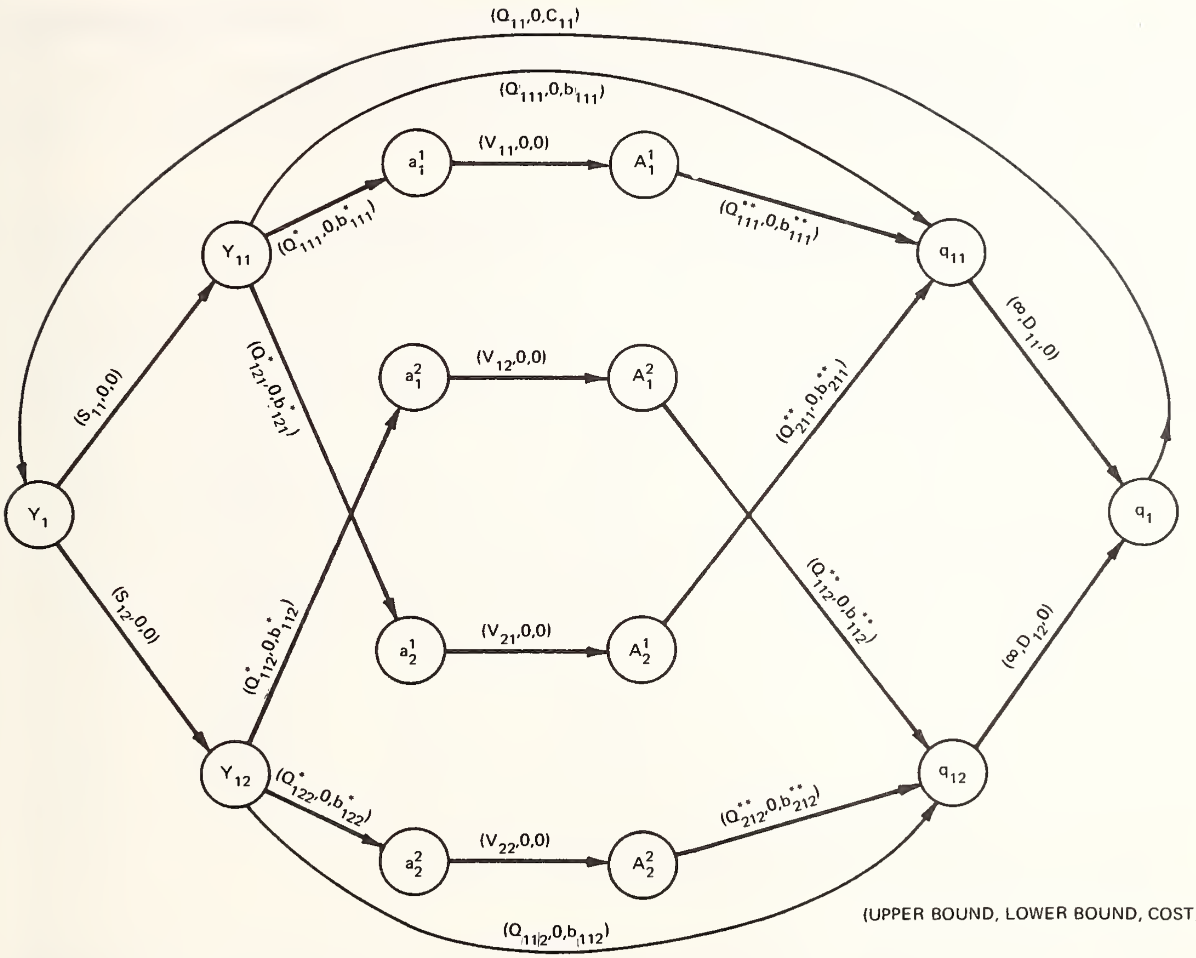


FIGURE 2. Single-source, single-sink, two-commodity extended truck assignment problem with two intermediate nodes

An example graph for $m=1$, $n=1$, $p=2$, with two intermediate nodes is shown in Figure 2.

COMPUTATIONAL RESULTS

Trial runs to compare the original Szwarc formulation with the (unextended) out-of-kilter form of the problem were made in FORTRAN IV on an IBM 7094. The out-of-kilter program used was IBM Share Distribution #3536, "Out of Kilter Network or Transportation Problem Solver." A program written by the authors incorporating a direct coding of the Ford and Fulkerson transportation algorithm [3] was used to solve the Szwarc form. In all cases in which demand equals supply, the efficiency of Szwarc's decomposition into several smaller transportation problems is such that it is always significantly faster than the out-of-kilter formulation. For this reason, trials were made only with cases in which demand is not equal to supply. Trial networks were generated randomly. Results are shown in Table 1, where the running times are averages of calculation time in three trials, not including generation of networks or output time. The results show that, although the Szwarc formulation is faster with a small number of commodities, the out-of-kilter formulation becomes more efficient as the number of com-

modities increases. This is quite predictable because the number of arcs in the Szwarc formulation is proportional to p^2, while the number of arcs in our formulation is proportional to p.

A few runs were also made with the out-of-kilter formulation for extended problems (with intermediate nodes). These results are shown in Table 2.

TABLE 1. *Comparison of Computation Times for Truck Assignment Problems (Unextended) Using the Szwarc and Out-of-Kilter Formulations*

Sources	Sinks	Commodities	Szwarc formulation		Out-of-kilter formulation	
			Size [a]	Time (sec)	Size [b]	Time (sec)
4	4	4	20x20	1.0	112x40	2.0
4	4	8	36x36	5.8	208x72	5.9
4	4	12	52x52	15.4	304x104	12.4
4	4	16	68x68	36.8	400x136	22.2
5	4	3	19x17	0.7	107x36	1.4
5	4	6	34x29	4.4	194x63	5.0
5	4	10	54x45	14.5	310x99	11.0
5	4	12	64x53	22.8	368x117	15.8
6	6	4	30x30	3.8	228x60	5.7
6	6	6	42x42	10.5	324x84	11.0
6	6	8	54x54	19.9	420x108	18.6
6	6	10	66x66	32.8	516x132	28.1

[a] (Number of sources) × (Number of sinks) in Hitchcock transportation problem.
[b] (Number of arcs) × (Number of nodes) in out-of-kilter problem.

TABLE 2. *Computation Time for Solving Some Multicommodity Trans-Shipment Problems Using the Out-of-Kilter Algorithm*

Number of sources	Number of int. nodes	Number of sinks	Number of commodities	Number of arcs	Number of nodes	Execution time (sec)
5	5	5	2	205	42	2.4
10	3	3	3	285	59	4.8
15	5	2	2	304	56	4.8
7	4	4	3	288	59	4.8

CONCLUSIONS

The Szwarc truck assignment problem has been extended to a more general transshipment network with upper bound capacity constraints on arcs by stating the problem in graph form and solving it by the out-of-kilter algorithm. In addition, lower bound capacities on arcs and on facilities might be established if needed without changing the nature of the problem. Szwarc's original problem may also be solved by this method. If the problem to be solved fits Szwarc's format and if either the total supply exactly equals the total demand and/or the number of commodities is small, then Szwarc's solution procedure is preferred. If there is a large number of commodities and the total supply exceeds the total demand, then our solution procedure is preferred.

ACKNOWLEDGMENTS

The Authors wish to acknowledge the support of the Department of Health, Education, and Welfare, Environmental Control Administration through grants 5T01 UI 01049 and 1R01 UI 00828.

REFERENCES

[1] Appelgren, Leif H., "A Column Generation Algorithm for a Ship Scheduling Problem," Transportation Science *3*, 53–68 (1969).

[2] Bellmore, M., G. Bennington, and S. Lubore, "Further Extensions of the Tanker Scheduling Problem," submitted to Transportation Science.

[3] Ford, L. R., Jr. and D. Fulkerson, *Flow in Networks* (Princeton University Press, Princeton, New Jersey, 1962).

[4] Fulkerson, D. R., "An Out of Kilter Method for Minimal Cost Flow Problems," J. Soc. Industrial and Applied Math. *9*, 18–27 (1961).

[5] Levin, Amos, *Some Fleet Routing and Scheduling Problems for Air Transportation Systems*, Flight Transportation Laboratory Report R–68–5, Massachusetts Institute of Technology (1969).

[6] Szwarc, W., "The Truck Assignment Problem," Nav. Res. Log. Quart. *14*, 529–557 (1967).

OPTIMUM POSITIONS FOR m AIRPORTS

Thomas L. Saaty

University of Pennsylvania

1. PROBLEM

There is a set of n factories, which are located throughout the United States. The wares produced by these factories are delivered to local airports, where they are flown to one of two overseas airports (Manila or Frankfurt), called distribution centers. The centers distribute the wares to a number of neighboring areas. Each center has a required tonnage to be delivered by air cargo.

The problem is to find the optimum location for a single airport in order to minimize the total transportation costs; it is also desired to find the optimum position for two airports, three airports, . . ., m airports by use of the minimum cost criterion.

The real situation requires the determination of subsets of one, two, three—up to m airports, which yield minimum shipping costs among a total of m airports already in existence.

Many papers have been written on this type of problem under various titles including "warehouse" and "location allocation" problem. A bibliography is included which may help the reader extend his knowledge of the subject.

Instead of a direct search for an exact numerical solution to this nonlinear problem we attempt to characterize the solution under various hypotheses relating to the geometry in order to bring out, whenever possible, some of its more interesting properties. This characterization provides useful insight to the decision-maker in developing a strategy as to which existing airports to keep operating, as to those whose potential capacity should be increased and as to which ones to shut down.

With respect to objectives, there are essentially two different ways to view this problem:

1. The first, or long-range plan, is to make an estimated aggregate of expected demand at each destination for a long period under varying circumstances of peace and war and determine positions for airports to satisfy this demand with minimum transportation costs.

2. The second, or short-range plan, is to determine positions for airports under peaceful requirements, but subject to probabilities that war might occur at either destination area. Local wars are assumed to happen with notable frequency. As we shall see, the first method leads to a clustering of airports on the east and west coasts of the continental United States, while the second leads to a distribution of airports over the country.

We assume that the total shipping capacity of the airports is not less than the requirements at the destinations.

2. AN ALGEBRAIC FORMULATION OF THE FIRST PROBLEM THE CASE OF A SINGLE AIRPORT

If we label the factories $i=1, 2, \ldots, n$ and assign them coordinates (x_i, y_i) and if C_i is the of shipping a single ton per mile from the ith factory from which S_i units $i=1, \ldots, n$ are shipped to a single airport and if C_{n+1}, C_{n+2} are the shipping costs per ton mile to the two destinations whose

coordinates are (x_{n+1}, y_{n+1}) and (x_{n+2}, y_{n+2}) and if the quantities shipped from the airport to the destinations are S_{n+1} and S_{n+2}, respectively. we wish to locate the coordinates (α, β) of the airport such that

$$f(\alpha, \beta) = \sum_{i=1}^{n+2} C_i S_i [(x_i - \alpha)^2 + (y_i - \beta)^2]^{1/2}$$

is a minimum subject to

$$\sum_{i=1}^{n} S_i = S_{n+1} + S_{n+2}$$

and to the fact that α, β should lie within the continental limits of the United States.

Differentiation with respect to α and β give, respectively,

$$\frac{\partial f}{\partial \alpha} = \sum_{i=1}^{n+2} C_i S_i \frac{x_i - \alpha}{[(x_i - \alpha)^2 + (y_i - \beta)^2]^{1/2}} = 0$$

and

$$\frac{\partial f}{\partial \beta} = \sum_{i=1}^{n+2} C_i S_i \frac{y_i - \beta}{[(x_i - \alpha)^2 + (y_i - \beta)^2]^{1/2}} = 0,$$

subject to

$$\sum_{i=1}^{n} S_i = S_{n+1} + S_{n+2}$$

constrained to lie in the continental United States.

The result is a complex nonlinear problem which cannot be solved in closed form and which would be considerably more complicated if several airport positions were to be considered where with each airport a certain capacity is associated.

The Case of *m* Airports

In the general case, we seek the positions of airports (α_j, β_j), $j=1, \ldots, m$, in the continental United States which minimize the shipping costs. If we let C_{ij} denote the shipping cost per ton-mile, and let S_{ij} be the number of tons shipped from factory i to airport j, $i=1, \ldots, n+2$; $j=1, \ldots, m$, we must minimize

$$\sum_{i,j=1}^{n+2;\, m} C_{ij} S_{ij} [(x_i - \alpha_j)^2 + (y_i - \beta_j)^2]^{1/2},$$

subject to

$$\sum_{i,j=1}^{n,m} S_{ij} = \sum_{j=1}^{m} S_{n+1,j} + S_{n+2,j}.$$

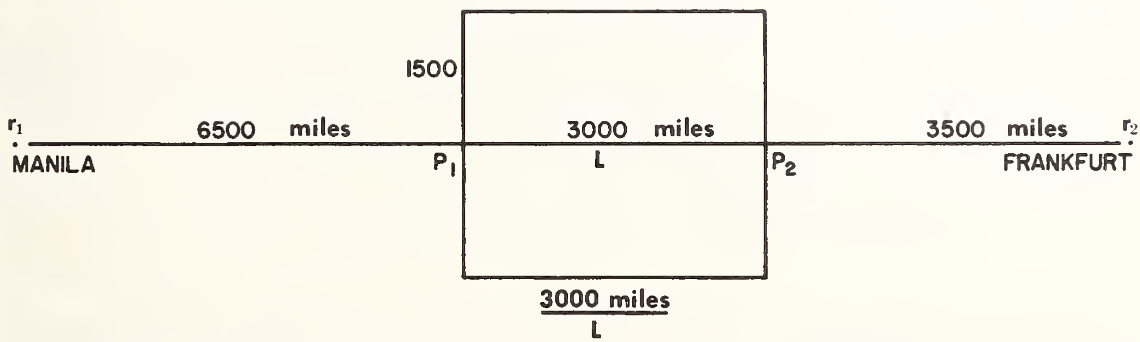
Since the algebra does not yield a useful picture of airport distribution, we turn to a geometric approach.

3. GEOMETRIC FORMULATION

We shall first examine the problem assuming the long range flow approach. Because of the difficulty of obtaining an explicit analytic answer using the above algebraic formulation, we need to make the following simplifying assumption:

> The cost of shipping by land from factories to airports is considerably less (perhaps by an order of magnitude) than shipping by air. Thus any material shipped from a factory to an airport on a coast would cost less by land than if it were to be flown there or directly to a destination.

We also assume (without loss of generality) that the two destinations lie on a straight line, L, which passes through the center of the United States with the west coast located at P_1 and the east coast at P_2 as in the following diagram.



Assume that the requirement at Manila is r_1 and that at Frankfurt is r_2 (instead of S_{n+1} and S_{n+2} previously used). And let C be the shipping cost per ton-mile by air. We also assume that the quantities shipped from the airports are equal to $r_1 + r_2$. We assume for the present that the factories do not enter into the problem and consider only the shipping costs from the airports.

LEMMA 1: If $r_1 = r_2$ then the optimum position for a single airport is anywhere on the line between P_1 and P_2.

PROOF: Let the airport be located at distance d_1 from Manilla and distance $d - d_1$ from Frankfurt, where d is the distance of Frankfurt from Manila. Then the total shipping cost is:

$$Cr_1 d_1 + Cr_2(d - d_1) = Cr_1 d,$$

and this is constant for all points between P_1 and P_2.

LEMMA 2: If $r_1 > r_2$, then the optimum position for a single airport is on the west coast at P_1.

PROOF: By use of Lemma 1, the airport may be located anywhere on L between P_1 and P_2; however, if we write $r_1 = r_2 + R$, then the greatest cost saving benefit for shipping the quantity R to Manila is obtained by placing the airport at P_1.

This follows from substituting for r_1 above. Then the total shipping cost is:

$$C(r_2+R)d_1+Cr_2(d-d_1)=Cr_2d+CRd_1,$$

which is minimized by choosing d_1 as small as possible.

In the case of two airports and equal requirements at both destinations, note that one airport must have a capacity equal to or exceeding the requirement at one of the destinations.

LEMMA 3: If $r_1=r_2$ then the optimum position for two airports A_1 and A_2 with capacities K_1 and K_2, respectively, with $K_1>K_2$ is to put one airport at P_1 and the other at P_2.

PROOF: Obviously, $K_1>r$ where r is the common value of r_1 and r_2. It follows that $K_2<r$. If d_1, d_2, and d are the distances of A_1, A_2, and Frankfurt from Manilla, respectively, then the shipping cost is given by

$$C(d-d_2)K_2+Cd_1r+C(d-d_1)(r-K_2)=Cdr-CK_2(d_2-d_1).$$

if and only if $d_1<d_2$. A symmetric argument applies when $d_2<d_1$. This expression is minimized with respect to d_1 and d_2 by making (d_2-d_1) as large as possible and hence by making d_2 as large as possible and d_1 as small as possible. This proves the lemma.

Consider two airports, A_1 and A_2 with capacities K_1 and K_2, respectively: let $r_1\geqslant r_2$ and suppose that $K_1\geqslant r_1$. It then follows from $K_1+K_2=r_1+r_2$ that $K_1\geqslant K_2$. Let $D_1=d_1$ be the distance from Manilla to A_1, $D_2=d_2-d_1$ the distance from A_1 to A_2 and $D_3=d-d_2$ the distance from A_2 to Frankfurt. The shipping cost is given by

$$C[D_1r_1+(D_2+D_3)(K_1-r_1)+D_3K_2]=C[D_1r_1+D_2(K_1-r_1)+D_3r_2].$$

Having used the fact that $K_1+K_2=r_1+r_2$.

Now the second and third terms of this expression say that the material K_1-r_1 may be shipped to A_2 and then a quantity r_2 is shipped from A_2.

By Lemma 2, the optimum position for A_1 between a destination requiring a quantity r_1 and one requiring $K_1-r_1<r_1$ is to put A_1 nearest to the first destination no matter where the second destination may be. Then, to minimize the shipping cost, A_2 should be placed as close to the destination requiring r_2 as possible; i.e., at P_2.

Again, suppose that $K_1<r_1$ and that $K_1\geqslant K_2$ (if the opposite holds, we exchange the airport labels); obviously we have $K_2>r_2$. The shipping cost is given by:

$$C[D_1K_1+(D_1+D_2)(r_1-K_1)+D_3r_2]=C[D_1r_1+D_2(r_1-K_1)+D_3r_2].$$

Thus the position of A_2 from which shipment is made to A_1 and to Frankfurt depends on whether $r_1-K_1\gtreqless r_2$.

We have proved:

LEMMA 4: If $r_1\geqslant r_2$ then the optimum position for two airports A_1 and A_2 with capacities K_1 and K_2, respectively, is obtained as follows:

If $K_1>r_1$ put A_1 at P_1 and A_2 at P_2.

$$\text{If } K_1 \leqslant r_1 \text{ put } A_1 \text{ at } P_1 \text{ and put } A_2 \text{ at } \begin{cases} P_1 \text{ if } r_1 - K_1 > r_2 \\ P_2 \text{ if } r_1 - K_1 \leqslant r_2. \end{cases}$$

The cost in the first case is given by $Cr_1d_1 + CK_2(d - d_2) + C(K_1 - r_1)(d - d_1)$ and in the second case by $Cr_1d_1 + Cr_2(d - d_1)$ or $CK_1d_1 + C(r_1 - K_1)d_2 + Cr_2(d - d_2)$.

REMARK: Strictly speaking, if $r_1 - K_1 = r_2$, then A_2 can be put any place on the line segment P_1P_2 (by Lemma 1). For exact equality, this would be piddling in a practical sense; however, since it is obviously desirable to have the freedom of placement anywhere on the continent, it might be useful to define a near equality. For example, if

$$|(r_1 - K_1) - r_2| < M$$

then we could consider $r_1 - K_1 = r_2$, and A_2 could be placed anywhere in P_1P_2. Here M would be that differential amount of material such that the cost of *not* placing the airport in the theoretically optimum position would be outweighed by the factors extraneous to this discussion.

THEOREM 1: Let $r_1 \geqslant r_2$, and let $A_1, \ldots A_m$ be airports with capacities $K_1, \ldots K_m$, respectively. The optimum position for the airports which yields the minimum transportation cost is obtained by dividing them into two subsets and taking the total capacity for each subset for all such possible pairs of partitions and evaluating the three cost expressions of Lemma 4 for each pair and noting the minimum of the three. That pair which yields the smallest minimum value gives the optimum solution.

PROOF: It is clear that every airport must lie either at P_1 or at P_2; otherwise, by Lemma 4 the transportation cost would be higher from this airport to the destinations. Since the airports are divided into two clusters, we may apply Lemma 4 to find the best position by taking the airports one against $(m-1)$, there are m such possible partitions; then by taking 2 against $(m-2)$, there are $\binom{m}{2}$ such possible partitions and so on.

REMARK: In practice, one attempts to group the airports into two clusters in such a way that the total capacities of clusters is as near to r_1 as possible, and hence it would follow that the total capacity of the second cluster is as near to r_2 as possible. Also, in practice, the capabilities of the airports have a certain built-in flexibility which permits expansion during crisis periods followed by contraction during peaceful interludes.

The emerging pattern from this discussion is the following:

THEOREM 2: If $r_1 \geqslant r_2$ the optimum position for m airports $A_1, \ldots, A_m$ with given capacities $K_1, K_2, \ldots, K_m$ where $K_1 \leqslant K_2 \leqslant \ldots \leqslant K_m$ is to cluster the airports at P_1 and P_2 in such a way that A_1 is at P_1 and if $r_1 - K_1 > K_2$ to put A_2 at P_1 and if $r_1 - (K_1 + K_2) > K_3$ to put A_3 at P_1 and so on until one obtains the reverse inequality in which case all remaining airports are located at P_2. Similar results may be derived for the case $r_2 > r_1$.

4. THE EFFECT OF FACTORIES

As already pointed out, it is difficult to use the factories without using the algebraic approach; however, it may be useful to make some simplifying assumptions just to see what sort of geometric patterns they suggest for airports.

If there is a factory, for example, on the west coast near P_1, it is possible that the position of the

airport would be moved up or down along the coast in order to optimize the shipping cost from the factory to the airport and from the airport to the destination.

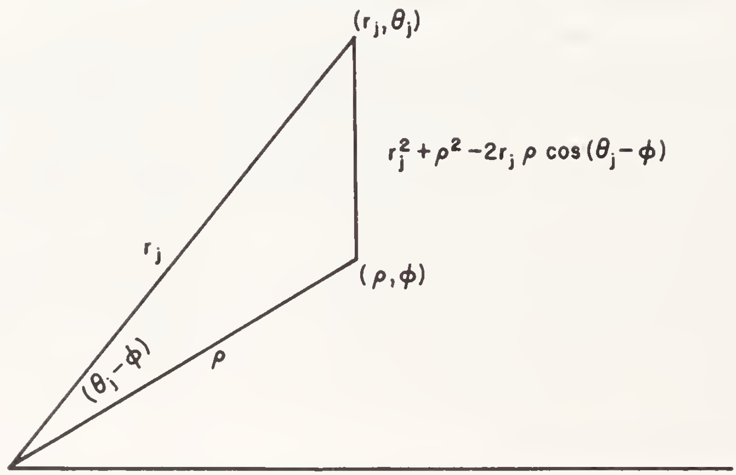
Assume that all factories ship to the airports a homogeneous product whose total volume equals the demand. Clearly, we have:

LEMMA 6: It is possible to divide the rectangle representing the United States by a north-south straight line such that all factories on each side of the line ship their homogeneous product to the airports on the coast of that side. Factories which lie on the line can ship in both directions to meet the demands.

PROOF: Trivial.

LEMMA 7: Given a single airport (ρ, φ) and a group of factories (r_j, θ_j), $j=1, \ldots, m$, all given in polar coordinates, then the optimal location of the airport must satisfy $\min_j \theta_j < \varphi < \max_j \theta_j$.

PROOF: If V_j denotes the capacity of the jth factory and C_j is its shipping cost per ton-mile to (ρ, φ), then the cost expression in polar form is given by



$$f(\rho, \varphi)=\rho C \sum_1^m V_j+\sum_1^m C_j V_j [r_j^2+\rho^2-2r_j\rho \cos(\theta_j-\varphi)]^{1/2}$$

$$\frac{\partial f}{\partial \varphi}=\sum_1^m -C_j V_j \frac{r_j \rho \sin(\theta_j-\varphi)}{[r_j^2+\rho^2-2r_j\rho\cos(\theta_j-\varphi)]^{1/2}} \qquad -\frac{\pi}{2} \leqslant \varphi \leqslant \frac{\pi}{2}$$

for $\varphi > \text{Max}\{\theta_j\}$, $\sin(\theta_j-\varphi) < 0$ for all j and hence $\frac{\partial f}{\partial \varphi} > 0$

for $\varphi < \text{Min}\{\theta_j\}$, $\sin(\theta_j-\varphi) > 0$ for all j and hence $\frac{\partial f}{\partial \varphi} < 0$

Thus, any φ for which $\frac{\partial f}{\partial \varphi}=0$, and there must be at least one, must lie in the range

$$\text{Min}\{\theta_j\} < \varphi < \text{Max}\{\theta_j\}$$

as required by the lemma.

LEMMA 8: Subject to the hypotheses of the previous lemma and a sufficiently high value of C, the optimum airport will be located on the coast.

PROOF:

$$\frac{\partial f}{\partial \rho}=C\sum_1^m V_j+\sum_1^m C_j V_j \frac{\rho-r_j\cos(\theta_j-\varphi)}{[r_j^2+\rho^2-2r_j\rho\cos(\theta_j-\varphi)]^{1/2}}.$$

If $\cos(\theta_j - \varphi)$ is replaced by unity, then $\frac{\partial f}{\partial \rho}$ is constant independent of ρ.

This value applies whenever ρ does not fall in the ocean; otherwise, the airport would be located on the coast.

In general, $\theta_j - \varphi$ would be small and hence $\cos(\theta_j - \varphi)$ would be close to unity. In that case, the ratio of the first term to the denominator gives a convex combination of the r_j and says that the airport lies in the convex hull of the factories.

However, this quantity is pulled to the left (towards the coast) or decreased according to the relative magnitude of the second term and its denominator. The higher the air costs, the more certain that the airport will lie on the coast.

Based on Lemmas 6 and 7, optimal airports can be set up on both coasts to channel material to the destinations. Some capacity will remain to be filled by factories located on or near the central line of Lemma 6. (If the capacity of the airports just matches that of the factories and the requirements, then more factories may be centrally channelled than the one factory of Lemma 6.) The factories on each side of the line which exhaust the requirements of a mixed central airport *and* have a minimum weighted center of gravity would then be channelled through one airport located by the one-airport problem solution method. That is, the number of factories which will be left on each side after exhaustion by an integral number of airports is computed (a simple task) and the particular factories are located by the single-airport method.

The second general approach to the problem is to assign probabilities to rising demand at either destination. In such a situation, it is clear that Lemma 1 would be modified in favor of the destinations with greater demand. However, the purpose in this case may be to locate the airport so as to equalize the costs of shipping to either destination. Considerations of this type lead to a spread of the airport clusters inland tending to occupy a balanced position in the case of sudden rising demand at either destination.

5. EXAMPLES OF THE EFFECT OF UNCERTAINTY

In the case of a mixed sequence of peace and war occurrences, we must define what the objective function is. Here, the expected value may not be a good measure. Thus, if war is a rare event, then it would not make sense to position the airport at an average position somewhere between where it would be under conditions of total peace and where its position comes out under conditions of total war. The reason for this is that war frequency and duration are both small; and, hence, one would incur a permanently high cost by shipping from this new position for the very long duration of peace requirements. Even if the frequency of war is high, its duration (not considered in the problem) may be short; and, hence, again, it does not seem advisable to solve the problem in a probabilistic setting unless the total duration of wars is considerable and needs to be taken into account.

Failure to give accurate estimates of the probability of war aggravates the usefulness of a stochastic formulation of the problem. It follows that unless one can extrapolate from the past on war frequency and duration, it would be more useful to apply the deterministic formulation for general decision-making as to the optimum location of airports.

Below we pursue our analysis under the assumptions that it is possible to estimate the probability of a war and that wars last long enough to require serious consideration.

Let us now examine the situation for two airports in which probabilities are attached to $r_1 \geqslant r_2$ and $r_2 \geqslant r_1$ and to $K_1 \geqslant r_1$ and $K_1 < r_1$. We have for our cost functions:

$$CP(r_1 \geqslant r_2)\ \{[D_1r_1 + D_2(K_1 - r_1) + D_3r_2]P(K_1 \geqslant r_1) + [D_1r_1 + D_2(r_1 - K_1) + D_3r_2]\ [1 - P(K_1 \geqslant r_1)]\}$$

$$+ C[1 - P(r_1 \geqslant r_2)]\ \{[D_1r_1 + D_2(K_1 - r_1) + D_3r_2]P(K_1 \geqslant r_1) + [D_1r_1 + D_2\,(r_1 - K_1) + D_3r_2][1$$

$$- P(K_1 \geqslant r_1)]\} = C\{D_1r_1 + D_3r_2 - [1 - 2P(K_1 \geqslant r_1)]D_2(K_1 - r_1)\} = C\{D_1[r_1 + \{1 - 2P(K_1 \geqslant r_1)\}(K_1$$

$$- r_1)] + D_3[r_2 + \{1 - 2P(K_1 \geqslant r_1)\}(K_1 - r_1)] - d[1 - 2P(K_1 \geqslant r_1)](K_1 - r_1)\}.$$

The problem is to find D_1 and D_3 which minimize this expression, subject to the constraints that they fall in the continental United States.

This is a stochastic linear programming problem.

The net result for a pattern of m airports would be (we conjecture) roughly as follows: A clustering on the coasts and elsewhere in the center (indicating that the probability of war in Europe is higher), otherwise close to the west coast if the probabilities are the same, since the distance from the west coast to Manila and to Frankfurt are about the same. Between these three clusters there is a thin scattering of airports to account for various values of the probability distribution of war and peace. These ideas can be formalized along similar lines to the foregoing.

In conclusion, we note that generally the problem of war cannot be handled without considering the defense and security of planes and airports. Such a constraint would diminish the size of the region in which the allocation is made.

6. ACKNOWLEDGMENTS

I am grateful to Leon Goldstein, Larry Bein, and Hani Mesak for helpful comments, and to an anonymous referee for his assistance with the bibliography.

BIBLIOGRAPHY

Cooper, L., "Heuristic Methods for Location-Allocation Problems," SIAM Review *6*, 37–52 (1964).

Cooper, L., "Location-Allocation Problems," Operations Research *11*, 331–343 (1963).

Cooper, L., "Solutions of Generalized Locational Equilibrium Models," Journal of Regional Science, *7*, 1–18 (1967).

Efroymson, M. A. and T. L. Ray, "A Branch Bound Algorithm for Plant Location," Operations Research, *14*, 361–368 (1966).

Friedrich, C. J., *Alfred Weber's Theory of the Location of Industries* (University of Chicago Press, Chicago, 1929).

Hardgrave, W. W., "Location-Allocation Problems: A Survey," Operations Research *16*, Supplement 1 (1968), p. B–84 (abstract).

Isard, W., *Location and Space-Economy* (Technology Press, M.I.T., Cambridge, 1956).

Jordan, R. H. (Project Leader), "Systems Analysis of Inland Consolidated Centers for the United States Maritime Administration," National Bureau of Standards Report 9892, NBS Project 4314422,

U.S. Department of Commerce (1969).

Kuhn, H. W. and R. E. Kuenne, "An Efficient Algorithm for the Numerical Solution of the Generalized Weber Problem in Spatial Economics," Journal of Regional Science, *4*, 21–23 (1962).

Miehle, W., "Link-Length Minimization in Networks," Operations Research *6*, 232–243 (1958).

Palermo, F. P., "A Network Minimization Problem," IBM Journal of Research and Development, *5*, 335–337 (1961).

INCREMENTAL APPROXIMATION OF OPTIMAL ALLOCATIONS*

Lawrence D. Stone

Daniel H. Wagner, Associates,
Paoli, Pennsylvania

ABSTRACT

This paper concerns the approximation of optimal allocations by Δ allocations. Δ allocations are obtained by fixing an increment Δ of effort and deciding at each step upon a single cell in which to allocate the entire increment. It is shown that Δ allocations may be used as a simple method of approximating optimal allocations of effort resulting from constrained separable optimization problems involving a finite number of cells. The results are applied to find Δ allocations (called Δ plans) which approximate optimal search plans. Δ plans have the property that as $\Delta \to 0$, the mean time to find the target using a Δ plan approaches the mean time when using the optimal plan. Δ plans have the advantage that they are easily computed and more easily realized in practice than optimal plans which tend to be difficult to calculate and to call for spreading impractically small amounts of effort over large areas.

1. INTRODUCTION

Before stating our results in a mathematical fashion, we discuss the motivation for studying methods of approximating optimal allocations. The motivation arises from the study of optimal search plans.

Optimal search plans have been found for a large class of searches for a stationary object. Koopman [4] and DeGuenin [1] have found optimal plans when the search sensor has perfect discrimination. In the case of sensors with uncertain sweep width, optimal plans are given in [5]. In [7] and [8], optimal plans have been found for several classes of searches involving false targets. Typically, optimal plans have a complicated form even when one is able to overcome the analytic difficulties involved in finding their explicit expressions. This complicated form makes it difficult to provide planning advice without the aid of computers. Even in the cases where the functional form of the optimal plan is simple (see, for example, p. 41 of [3]), the plans typically call for spreading small amounts of search over ever expanding areas; however, this is very difficult to do in practice.

As a method of overcoming some of the difficulties involved in using optimal plans, we present a class of plans called Δ plans. These plans are executed in a step by step fashion. At each step one is required to calculate simple ratios and allocate Δ amount of effort to a single cell. These plans can be found by using a desk calculator and do not require that small amounts of search effort be spread over large areas. Furthermore, these plans approximate the optimal plan in the sense that as $\Delta \to 0$, the mean time to find the target using a Δ plan approaches the mean time using the optimal plan.

One major restriction on using Δ plans is that they apply only when the target location distribution is specified by a finite number of cells R_j, $1 \leq j \leq J$, such that the target is in cell R_j with probability

*This research was supported by the Naval Analysis Program (Code 462), Office of Naval Research under Contract No. N 00014–69–C–0435.

p_j. For most search operations, however, the target location distribution may be specified in this manner (see [6]). In mathematical terms our results may be described as follows.

For functions, f, of two variables we shall use $f(x,\cdot)$ to indicate the function obtained from f by fixing the first variable at x. Similarly, when fixing the second variable at y, we write $f(\cdot, y)$. Let $\mathbf{J}$ be a countable set of indices which for convenience we take to be a subset of the positive integers. For $j\epsilon\mathbf{J}$ let $e(j,\cdot)$ *and* $c(j,\cdot)$ be nonnegative functions defined on $[0,\infty)$ such that $e(j,0)=c(j,0)=0$. Let G be the set of nonnegative functions g defined on $\mathbf{J}$ and let

$$E(g)=\Sigma e(j,g(j)) \qquad \text{and} \qquad C(g)=\Sigma c(j,g(j)),$$

where summations without indicated ranges are understood to run over all of $\mathbf{J}$. Let J denote the cardinality of $\mathbf{J}$.

An *allocation* is a function q: $\mathbf{J}\times[0,\infty)\to[0,\infty)$ such that

(i) $q(j,\cdot)$ is increasing and continuous for $j\epsilon\mathbf{J}$

(ii) $\Sigma\, q(j,s)=s \qquad$ for $s\geqslant 0$.

We refer to $q(j, s)$ as the amount of effort in the jth cell, R_j, to $c(j. q(, s))$ as the cost of that effort and to $e(j,q(j,s))$ as the effectiveness resulting from the effort. Let Q be the set of all allocations q. Thus $C(q(\cdot,s))$ and $E(q(\cdot,s))$ are the global cost and effectiveness resulting from $q(\cdot,s)$. Define

$$\mu(q)=\int_0^\infty C(q(\cdot,s))dE(q(\cdot,s)),$$

whenever the integral exists in the Stieltjes sense. Then q^* is an *optimal* allocation if

$$\mu(q^*)\leqslant\mu(q) \qquad \text{for } q\epsilon Q.$$

We shall refer to $\mu(q)$ as the *mean cost* of the allocation q.

If f is a real valued function of one or more variables, we let f' denote the partial derivative of f with respect to the last variable. In Section 2, Theorem 2.1 finds the optimal allocation q^* under the conditions that for $j\epsilon\mathbf{J}$

(a) $e(j,\cdot)\leqslant B(j)$ and $\Sigma B(j)<\infty \qquad$ for some B:$\mathbf{J}\to[0,\infty)$.

(b) $e'(j,\cdot)$ and $c'(j,\cdot)$, are continuous and $c'(j,\cdot)>k \qquad$ for some $k>0$.

(c) $r_j=e'(j,\cdot)/c'(j,\cdot)$ is positive, continuous, and strictly decreasing.

The notion of a Δ allocation is defined in Section 3. Let μ_Δ be the mean cost resulting from a Δ allocation. Then Theorem 3.1 shows that if (1.1) is satisfied, $J<\infty$ and $c'(j,\cdot)\leqslant K$ for $j\epsilon\mathbf{J}$, then

$$\mu_\Delta\leqslant\mu(q^*)+2(J+1)K\Delta. \tag{1.2}$$

In Section 4 we show how the results of Section 3 may be used to approximate optimal search plans by Δ allocations which are called Δ plans in this special case. Section 5 demonstrates by example that the bound in (1.2) has the correct order of magnitude in the sense that there are Δ plans such that $\mu_\Delta-\mu(q^*)$ goes linearly to infinity as $\Delta\to\infty$.

2. OPTIMAL ALLOCATION

In Theorem 2.1 of this section, we find the optimal allocation q^* under the conditions given by (1.1). Theorems 2.1 and 2.2 are generalizations of Theorem 2 of [7] to allow more general cost functions, c. Observe that if (1.1) holds, then $\mu(q)$ is well defined for all $q \epsilon Q$. In addition we may define r_j^{-1} to be the inverse function for r_j and extend its domain to $(0, \infty)$ by letting $r_j^{-1}(z)=0$ for $z > r_j(0)$ and $r_j^{-1}(z)=\infty$ for $0 < z \leqslant \lim_{y\to\infty} r_j(y)=\beta_j$.

THEOREM 2.1: Assume conditions (1.1) are satisfied. Define $A(\gamma)=\sum r_j^{-1}(\gamma)$ for $\gamma > 0$. Then λ, the inverse function for A, is well defined on $(0, \infty)$. If

$$q^*(j, s) = r_j^{-1}(\gamma(s)) \qquad \text{for } s > 0,\ j \epsilon \mathbf{J}, \tag{2.1}$$

then $q^* \epsilon Q$,

$$E(g) \geqslant E(q^*(\cdot, s)) \text{ implies } C(g) \geqslant C(q^*(\cdot, s)) \qquad \text{for } g \epsilon G \tag{2.2}$$

and

$$\mu(q^*) \leqslant \mu(q) \qquad \text{for any } q \epsilon Q. \tag{2.3}$$

PROOF: Observe that

$$\frac{B(j)}{k} > \int_0^z \frac{e'(j, y)dy}{k} \geqslant \int_0^z \frac{e'(j, y)}{c'(j, y)}\,dy \geqslant z r_j(z) \qquad \text{for } z \geqslant 0.$$

Thus

$$r_j(z) \leqslant \frac{\mathrm{B}(j)}{k}\frac{1}{z}, \qquad \text{and} \qquad r_j^{-1}(\gamma) \leqslant \frac{\mathrm{B}(j)}{k}\frac{1}{\gamma} \qquad \text{for } \gamma > \beta_j.$$

Now we may write, for $\gamma > \max\{\beta_j : j \epsilon \mathbf{J}\}$,

$$A(\gamma) = \sum_{[j: r_j^{-1}(\gamma) > 0]} r_j^{-1}(\gamma) \leqslant \frac{1}{k\gamma} \sum B(j) < \infty. \tag{2.4}$$

By using the monotone convergence theorem, one may show that $\lim_{\gamma\to\infty} A(\gamma)=0, \lim_{\gamma\to 0+} A(\gamma)=\infty$ and that A is continuous. Furthermore A is strictly decreasing on the domain (Λ_l, Λ_u) where

$$\Lambda_l = \sup\{\Lambda : A(\Lambda)=\infty\} \qquad \Lambda_u = \inf\{\Lambda : A(\Lambda)=0\}.$$

Thus there is a unique function $\lambda: (0, \infty) \to (\Lambda_l, \Lambda_u)$ such that $A(\lambda(s)) = s$ for $s > 0$. Moreover, λ is continuous and strictly decreasing.

Set $q^*(j, 0) = 0$. Then $q^*(j, \cdot)$ is continuous and increasing and $\sum q^*(j, s) = s$ for $s > 0$. Hence $q^* \epsilon Q$. One may check that for each $s > 0$, q^* satisfies

$$e'(j, z) \geq \lambda(s)c'(j, z) \qquad 0 < z < q^*(j, s)$$

$$e'(j, z) \leq \lambda(s)c'(j, z) \qquad q^*(j, s) < z < \infty, \qquad \text{for } j \epsilon \mathbf{J}.$$

Thus $q^*(\cdot, s)$ satisfies the form of the Neyman-Pearson conditions given in [9], and by Theorem 1 and Remark 4 of [9], q^* satisfies (2.2) for each $s > 0$. The proof that (2.3) holds now follows in exactly the same manner as that given in the proof of Theorem 2 of [7]. This proves the theorem.

While the form of Theorem 2.1 given above is most convenient for our use, the theorem remains true in greater generality. In particular suppose that X is a topological space and that ν is a nonnegative finite measure defined on the Borel subsets $\mathscr{S}$ of X. Let e and c be Borel functions defined on $X \times (0, \infty)$, such that $e(x, 0) = c(x, 0) = 0$ for $x \epsilon X$. Then for nonnegative Borel functions g defined on X, we let

$$E(g) = \int_X e(x, g(x))d\nu(x) \qquad \text{and} \qquad C(f) = \int_X c(x, g(x))d\nu(x).$$

A simple modification of the proof of Theorem 2.1 yields the following theorem.

THEOREM 2.2: Let $|e| \leq B < \infty$. For $x \epsilon X$, let $e(x, \cdot)$ and $c(x, \cdot)$ be absolutely continuous and $r(x, \cdot) = e'(x, \cdot)/c'(x, \cdot)$ be positive, continuous and strictly decreasing. Let $r^{-1}(x, \cdot)$ be the inverse function for $r(x, \cdot)$ and extend the domain of $r^{-1}(x, \cdot)$ to $(0, \infty)$ by defining $r^{-1}(x, z) = 0$ for $z > r(x, 0)$ and $r^{-1}(x, z) = \infty$ for $z < \lim_{y \to \infty} r(x, y)$. If $c'(x, \cdot) > k$ for $x \epsilon X$, and some $k > 0$, then there exists a decreasing function λ defined on $(0, \infty)$ such that

$$q^*(x, s) = r^{-1}(x, \lambda(s)) \qquad \text{for } s > 0,\ x \epsilon X$$

satisfies

$$\int_X q^*(x, s)d\nu(x) = s,$$

and for any nonnegative Borel function g,

$$E(g) \geq E(q^*(\cdot, s)) \text{ implies } C(g) \geq C(q^*(\cdot, s)).$$

3. APPROXIMATION BY Δ ALLOCATIONS

We now define Δ allocations and show that they approximate the optimal allocation in mean cost. The advantage of a Δ allocation is that it calls for fixing an increment Δ and at each step in the plan allocating Δ amount of effort to only one cell. In contrast, the optimal allocation calls for spreading smaller and smaller amounts of effort over larger and larger areas. Thus, a Δ allocation is more likely to be operationally feasible.

To retain generality, we use effort as an undefined term. Possible definitions of effort include time, man-hours, money, track length, etc. In order to perform a Δ allocation, one fixes a positive number Δ. At each step, one allocates Δ amount of effort (in some fixed units) to a single cell as follows: Calculate

$$r_j(z_j) = \frac{e'(j,z_j)}{c'(j,z_j)} \qquad \text{for } j\epsilon\mathbf{J},$$

where z_j is the amount of effort placed in R_j. Allocate the next Δ amount of effort to the cell R_j having the highest value of $r_j(z_j)$. If there is more than one cell with the highest value, then one may allocate Δ to any one of these highest cells; one possible procedure is to choose the cell with the lowest index. Any plan generated in the above manner is called a Δ allocation. The mean cost of such an allocation is denoted μ_Δ. Fix Δ and a Δ allocation. Let $h(j,n)$ be the effort placed in R_j after n steps of the Δ allocation. In order to show that (1.2) holds we prove the following lemma.

LEMMA 3.1: Suppose the conditions of (1.1) are satisfied and that $J < \infty$. Then

$$\text{(3.1)} \qquad q^*(j,(n-J)\Delta) \leq h(j,n) \leq q^*(j,n\Delta) + \Delta, \qquad \text{for } j\epsilon\mathbf{J}$$

where we define $q^*(j, (n-J)\Delta) = 0$ when $n < J$.

PROOF: Let $\lambda_n^+ = \max r_j(h(j,n))$ where max without an indicated range is understood to run over all of $\mathbf{J}$. Then we claim

$$\text{(3.2)} \qquad r_j^{-1}(\lambda_n^+) \leq h(j,n) \leq r_j^{-1}(\lambda_n^+) + \Delta, \qquad j\epsilon\mathbf{J}.$$

To see that the left-hand side of (3.2) holds we observe that $\lambda_n^+ \geq r_j(h(j,n))$ by definition and then apply r_j^{-1} to both sides of this inequality. Since r_j^{-1} is decreasing, the left-hand side of (3.2) follows.

If $h(j,n) = 0$, we observe that the righthand side of (3.2) holds trivially. Suppose $h(j,n) = i\Delta$ for some $i \geq 1$. It follows that

$$\text{(3.3)} \qquad r_j((i-1)\Delta) \geq \lambda_n^+;$$

for if (3.3) does not hold, the last increment Δ placed in R_j was not added according to a Δ allocation. That is effort was added in R_j at a time when $r_j((i-1)\Delta)$ was not the highest ratio.

Thus

$$h(j, n) = (i-1)\Delta + \Delta \leq r_j^{-1}(\lambda_n^+) + \Delta$$

and (3.2) holds.

Summing (3.2) we obtain

$$\text{(3.4)} \qquad \Sigma r_j^{-1}(\lambda_n^+) \leq n\Delta \leq \Sigma r_j^{-1}(\lambda_n^+) + J\Delta.$$

Hence $(n-J)\Delta \leq \Sigma r_j^{-1}(\lambda_n^+)$. Letting $s = (n-J)\Delta$, we have by Theorem 2.1 that

$$q^*(j, (n-J)\Delta) = r_j^{-1}(\lambda(s)) \qquad \text{and} \qquad \Sigma r_j^{-1}(\lambda(s)) = (n-J)\Delta.$$

Since $\Sigma r_j^{-1}(\lambda(s)) \leq \Sigma r_j^{-1}(\lambda_n^+)$, it follows that $\lambda_n^+ \leq \lambda(s)$ and $q^*(j, (n-J)\Delta) = r_j^{-1}(\lambda(s)) \leq r_j^{-1}(\lambda_n^+)$. Thus we have shown the left-hand inequality in (3.1). By (3.4) and a similar argument, $\Sigma r_j^{-1}(\lambda_n^+) \leq n\Delta$ and $r_j^{-1}(\lambda_n^+) \leq q^*(j, n\Delta)$. The righthand side of (3.1) now follows from (3.2), and the lemma is proved.

Since a Δ allocation specifies the allocation of effort only at integer multiples of Δ, it is convenient to choose an allocation $q_\Delta \epsilon Q$, such that

$$q_\Delta(j, n\Delta) = h(j, n), \qquad \text{for } j \epsilon \mathbf{J}, \ n = 0, 1, 2, \ldots .$$

Such an allocation q_Δ clearly exists. Let

$$E_\Delta(s) = E(q_\Delta(\cdot, s)), \ C_\Delta(s) = C(q_\Delta(\cdot, s)),$$

and

$$E^*(s) = E(q^*(\cdot, s)), \ C^*(s) = C(q^*(\cdot, s)).$$

We now find a bound on the mean cost of using a Δ allocation.

THEOREM 3.2: Suppose that the conditions of (1.1) are satisfied, that $J < \infty$ and that $c'(j, \cdot) \leqslant K$ for $j \epsilon \mathbf{J}$. Then for any Δ allocation

$$\mu_\Delta \leqslant \mu(q^*) + 2(J+1)K\Delta. \tag{3.5}$$

PROOF: From (3.1) we obtain

$$E_\Delta(n\Delta) \geqslant E^*((n-J)\Delta) \qquad \text{for } n \geqslant 1,$$

where for convenience we define $E^*(s) = E_\Delta(s) = 0$ for $s \leqslant 0$. Thus for $(n-1)\Delta \leqslant s \leqslant n\Delta$,

$$E_\Delta(s) \geqslant E^*((n-J-1)\Delta) \geqslant E^*(s-(J+1)\Delta),$$

and

$$E_\Delta(s) \geqslant E^*(s-(J+1)\Delta) \qquad \text{for } s \geqslant 0. \tag{3.6}$$

Using the other half of (3.1) we obtain

$$\begin{aligned} C_\Delta(n\Delta) &\leqslant \Sigma c(j, q^*(j, n\Delta) + \Delta) \\ &\leqslant C(q^*(\cdot, n\Delta)) + JK\Delta. \end{aligned}$$

Thus for $(n-1)\Delta \leqslant s \leqslant n\Delta$,

$$\begin{aligned} C_\Delta(s) &\leqslant C^*(n\Delta) + JK\Delta \\ &\leqslant C^*((n-1)\Delta) + (J+1)K\Delta \\ &\leqslant C^*(s) + (J+1)K\Delta, \end{aligned}$$

and

$$C_\Delta(s) \leqslant C^*(s) + (J+1)K\Delta \qquad \text{for } s \geqslant 0. \tag{3.7}$$

Since E_Δ is monotone and continuous and C_Δ is monotone, one may use integration by parts and an argument similar to the one given to prove Lemma 1 on page 148 of [2] to verify that

$$\mu_\Delta = \int_0^\infty [1 - E_\Delta(s)] dC_\Delta(s). \tag{3.8}$$

Thus by (3.8) and (3.6)

$$\mu_\Delta \leq \int_0^\infty [1-E^*(s-(J+1)\Delta)]dC_\Delta(s). \tag{3.9}$$

An argument similar to the one yielding (3.8) combined with (3.7) gives

$$\int_0^\infty [1-E^*(s-(J+1)\Delta)]dC_\Delta(s)$$

$$= \int_0^\infty C_\Delta(s)dE^*(s-(J+1)\Delta) \leq (J+1)K\Delta + \int_0^\infty C^*(s)dE^*(s-(J+1)\Delta). \tag{3.10}$$

Since

$$C^*(s) \leq C^*(s-(J+1)\Delta)+(J+1)K\Delta,$$

(3.10) yields

$$\mu_\Delta \leq \mu(q^*)+2(J+1)K\Delta$$

and the theorem is proved.

For the special case where $c(j,z)=z$, we prove the following result which is stronger than Theorem 3.2.

THEOREM 3.3: Suppose $J<\infty$ and that for $j\in \mathbf{J}$, $c(j,z)=z$ for $z\geq 0$ and $e'(j,\cdot)$ is continuous and strictly decreasing. Then

$$\mu_\Delta \leq \mu(q^*)+(J+1)\Delta. \tag{3.11}$$

PROOF: The proof proceeds in the same manner as that of Theorem 3.2 to obtain (3.6). Since $c(j,z)=z$, we have $C_\Delta(s) = C^*(s)$ and

$$\mu_\Delta = \int_0^\infty [1-E_\Delta(s)]ds \leq \int_0^\infty [1-E^*(s-(J+1)\Delta)]ds$$

which proves the Theorem.

4. APPROXIMATION OF OPTIMAL SEARCH PLANS

In this section we show how Δ allocations may be used to approximate optimal search plans in the case where the search region consists of a finite number of cells R_j, $j=1, 2, \ldots, J$. The jth cell has probability p_j of containing the target and $\Sigma p_j=1$.

In order to make the search models of [7] or [8] fit into the framework developed in this paper, one must consider target location distributions which can be divided into a finite number of cells such that the density f of the distribution is constant over each cell. In many operations, the target location distribution is given in this manner (see [6]).

In order to make this paper self-contained, we briefly present the search model of [7] in a form which fits into the situation considered in this paper.

We consider search for a stationary target. The search may be complicated by the possibility of detecting false targets (i.e., objects which are not the target but cause a sensor reponse which cannot be distinguished from that of the target without further investigation). When a false target or a real target is detected, it becomes a *contact*.

The search takes place in two phases. The *broad search* phase is conducted using a sensor which can detect the target but not positively identify it. In order to investigate a contact, the broad search must stop and a *contact investigation* must begin. Once a contact investigation has begun, it must continue until the contact is identified. This is called *uninterrupted contact investigation*. At the end of a random time, the contact is correctly identified either as being or not being the target. We assume that investigation of one contact makes no contribution to investigation of any other contact or to the broad search. Also, once a contact has been investigated it is, in effect, eliminated and will not be classified as a contact if detected again.

We distinguish between two types of search time. Cumulative broad search time will be denoted by s. Cumulative time spent in all aspects of search and investigation will be denoted by t. To avoid confusion, we say a target (real or false) is *contacted* when it appears as a contact, and that it has been *identified* when contact investigation shows that contact to be a real or false target. We say the target has been *found* when it has been contacted and identified.

For the broad search it is assumed that for each cell R_j there is a *local effectiveness function* b_j defined on $[0, \infty)$ such that

(4.1)
(i) $0 \leq b_j \leq 1,\ b_j(0)=0,$ and $\lim_{z\to\infty} b_j(z)=1$

(ii) b_j', the derivative of b_j, exists, is strictly positive, continuous, and strictly decreasing.

If z amount of time is spent broad searching in R_j, then $b_j(z)$ is the probability of detecting the target given it is located in R_j. The assumption that b_j depends only on the amount of search time carries with it the implicit assumption that effort is applied uniformly over R_j.

For each cell R_j, we suppose that there is a $\delta_j \geq 0$, such that

$$Pr\left\{\begin{array}{c}\text{detecting exactly } k \text{ false targets in } R_j\\ \text{in } z \text{ amount of time spent broad}\\ \text{searching in } R_j\end{array}\right\} = e^{-\delta_j b_j(z)}\frac{[\delta_j b_j(z)]^k}{k!} \qquad \text{for } k=0, 1, \ldots .$$

Moreover, we assume that the mean time required to investigate a contact found in R_j is $T_j < \infty$.

If we take $\delta_j=0$ for $j\epsilon\mathbf{J}$, then the above model reduces to a discrete search region version of the model given by DeGuenin [1] with the exception that DeGuenin erroneously omitted the assumption that b_j be continuous.

A *search plan* is an allocation of a broad search time and a method of identifying contacts. For searches involving uninterrupted contact investigation we shall always assume that contact investigation is immediate. Thus the allocation of broad search completely specifies a search plan when using immediate and uninterrupted contact investigation. One may show by the same argument as that given in Section 3 of [7], that immediate contact investigation coupled with the q^* resulting from the definitions of e and c given in (4.2) below produces the smallest mean time to find the target when contact investigation is uninterrupted.

Define

$$e(j,z) = p_j b_j(z)$$

(4.2)

$$c(j,z) = z + T_j \delta_j b_j(z) \qquad j \in \mathbf{J}, \qquad z \geqslant 0.$$

Then $E(q(\cdot, s))$ is the probability of detecting the target by broad search time s using plan q and $C(q(\cdot, s))$ is the expected amount of time spent in broad search and contact investigation by broad search time s given the target has not been detected. Finally $\mu(q)$ is the mean time to detect the target using plan q. Since the contact which is the target is investigated immediately in all plans, minimizing μ is equivalent to minimizing the mean time to find the target.

Since e and c defined in (4.2) satisfy the conditions of Theorem 2.1, q^* is given by (2.1) where r_j^{-1} is the suitably extended inverse of

$$r_j(z) = \frac{p_j b_j'(z)}{1 + T_j \delta_j b_j'(z)} \qquad \text{for } z > 0.$$

Define a Δ plan as follows:

Allocate each increment of Δ units of broad search time to the cell having the highest value of $r_j(z_j)$, *where* z_j *is the amount of broad search time previously placed in* R_j *and investigate the resulting contacts until they are identified.*

COROLLARY 4.1: If $J < \infty$ and

$$\max \{1 + T_j \delta_j b_j'(0)\} = K < \infty$$

then

(4.3) $$\mu_\Delta \leqslant \mu^* + 2(J+1)K\Delta.$$

PROOF: The corollary follows directly from Theorem 3.2.

A similar although more complicated method may be used to approximate the optimal search plans generated in Section 4.1 of [8]. These Δ plans would call for adding Δ amount of broad search to the cell R_j having the highest value of

$$r_j(z_j, w_j) = \frac{p_j b_j'(z_j) a_j(w_j)}{1 + \delta_j b_j'(z_j) \alpha_j(w_j)},$$

where a_j and α_j are defined as the obvious analogs of a and α in [8], z_j gives the amount of time spent broad searching in R_j and w_j gives the amount of investigation time one is willing to devote to each contact generated in R_j. Suppose that j_o is the index of the cell to which the search is added. One then finds w_o such that

$$r_{j_o}(z_{j_o} + \Delta, w_o) = \frac{p_j a_j'(w_o)}{\delta[1 - A_j(w_o)]},$$

where A_j is defined as the analog of A in [8]. One is then willing to devote up to w_o total amount of contact investigation effort to each contact in R_j. Thus w_j becomes a function of z_j. We denote this by writing $w_j = v_j(z_j)$. If

$$\infty > K > 1 + \delta_j b_j'(z_j)[1 - A_j(v_j(z))] v_j'(z) \qquad \text{for } j \in \mathbf{J} \text{ and } z \geqslant 0,$$

then (4.3) holds.

5. EXAMPLE

When there are no false targets (i.e., $\delta_j=0$ for $j\epsilon\mathbf{J}$) Theorem 3.3 says that the penalty, in mean time, resulting from using a Δ plan is bounded by $(J+1)\Delta$. Obviously, as J or Δ approach infinity, this bound also approaches infinity. Since the bounds in the above analysis are somewhat rough, one might wonder if it is really possible for Δ plans to incur penalties which are linear in J and Δ as these quantities approach infinity. The example below answers this question in the affirmative by showing a case in which the penalty approaches infinity at the same rate as $(J\Delta)/2$.

Consider the situation where the target location distribution is uniform over a square of unit area. We suppose that this square is subdivided into J rectangles, each having equal area. We enumerate these rectangles in some order and let R_j be the jth rectangle. Thus, $p_j=1/J$, $1\leqslant j\leqslant J$. For $1\leqslant j\leqslant J$, let

$$\begin{aligned} b_j(z)&=1-e^{-Jz} \qquad \text{for } z\geqslant 0\\ \delta_j&=0. \end{aligned}$$

(Note that all times are assumed to be given in terms of a fixed time unit.) One may check that $\mu^*=1$ for this example.

We now consider a Δ plan which allocates the $(nJ+j)$th increment of broad search time to R_j for $n\geqslant 0$. Let μ_j be the mean time to find the target given the target is located in R_j. Consider the 1st rectangle. The mean time required to find the target given it is found during the first increment Δ of effort is

$$\frac{1-(1+J\Delta)e^{-J\Delta}}{J(1-e^{-J\Delta})}.$$

Thus

$$\mu_1=\frac{1-(1+J\Delta)e^{-J\Delta}}{J}+e^{-J\Delta}(J\Delta+\mu_1).$$

Solving for μ_1, we find

$$\mu_1=\frac{1}{J}+\frac{(J-1)\Delta e^{-J\Delta}}{(1-e^{-J\Delta})}.$$

It is easily seen that

$$\mu_j=(j-1)\Delta+\mu_1,\ j=1,\ .\ .\ .,J,$$

and

$$\mu_\Delta=\frac{1}{J}\sum_{j=1}^{J}\mu_j=\mu_1+\frac{(J-1)\Delta}{2}=\frac{1}{J}+\frac{(J-1)\Delta(1+e^{-J\Delta})}{2(1-e^{-J\Delta})}.$$

Since $\mu^*=1$,

$$\mu_\Delta=\mu^*+\frac{(J-1)\Delta}{2}\left[\frac{(1+e^{-J\Delta})}{(1-e^{-J\Delta})}-\frac{2}{J\Delta}\right].$$

Thus, the penalty resulting from using a Δ plan approaches infinity at the same rate as $J\Delta/2$ as J or Δ approaches infinity.

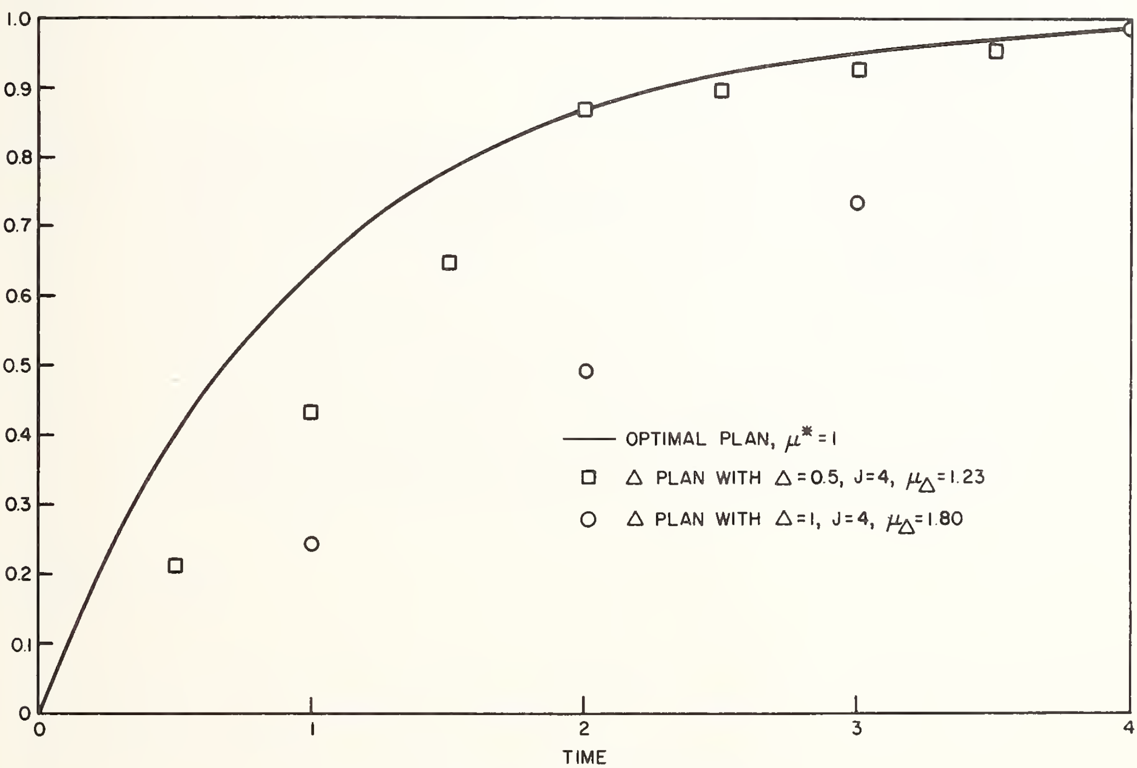


FIGURE 1. Probability of detection

In Figure 1, we have plotted the detection probabilities resulting from two of the Δ plans described above in the case where the search region is divided into four subregions. These probabilities are compared with the probabilities resulting from the optimal plan. For the first Δ plan the increment $\Delta=0.5$ has been chosen. Thus, the first 0.5 units of time are spent searching in R_1. The resulting probability of detection is 0.22. Note that the probability of detection for this plan and the optimal plan coincide at times 2 and 4. In most searches, however, the detection probability resulting from a Δ plan will always be strictly less than optimal. The second Δ plan results from taking $\Delta=1$, and as one would expect, it produces uniformly lower detection probabilities than either the optimal plan or the Δ plan with $\Delta=0.5$.

REFERENCES

[1] DeGuenin, J., "Optimum Distribution of Effort: An Extension of the Koopman Basic Theory," Operations Res. 9, 1–7 (1961).

[2] Feller, W., *An Introduction to Probability Theory and Its Application* (John Wiley & Sons, Inc., N.Y., 1966), Vol. II.

[3] Koopman, B. O., *Search and Screening*, OEG Report (1946).

[4] Koopman, B. O., "The Theory of Search III: The Optimum Distribution of Searching Effort," Operations Res., *5*, 613–626 (1957).

[5] Richardson, H. R. and Belkin, B., "Optimal Search with Uncertain Sweep Width," Operations Res., to appear.

[6] Richardson, H. R. and Stone, L. D., "Operations Analysis in the Underwater Search for the Scorpion," Nav. Res. Log. Quart., *18*, 141–157 (1971).

[7] Stone, L. D. and Stanshine, J. A., "Optimal Search Using Uninterrupted Contact Investigation," SIAM J. Appl. Math. *20*, 241–263 (1971).

[8] Stone, L. D., Stanshine, J. A., and Persinger, C. A., "Optimal Search in the Presence of Poisson-Distributed False Targets," to appear SIAM J. Appl. Math. *23* (July 1972).

[9] Wagner, D. H., "Nonlinear Functional Versions of the Neyman-Pearson Lemma," SIAM Review, *11*, 52–65 (1969).

THE PAYMENT SCHEDULING PROBLEM†

Richard C. Grinold

Center for Research in Management Science
University of California
Berkeley

ABSTRACT

Large complicated projects with interdependent activities can be described by project networks. Arcs represent activities, nodes represent events, and the network's structure defines the relation between activities and events. A schedule associates an occurrence time with each event; the project can be scheduled in several different ways. We assume that a known amount of cash changes hands at each event. Given any schedule the present value of all cash transactions can be calculated. The payment scheduling problem looks for a schedule that maximizes the present value of all transactions.

This problem was first introduced by Russell [2]; it is a nonlinear program with linear constraints and a nonconcave objective. This paper demonstrates that the payment scheduling problem can be transformed into an equivalent linear program. The linear program has the structure of a weighted distribution problem, and an efficient procedure is presented for its solution. The algorithm requires the solution of triangular systems of equations with all matrix coefficients equal to ± 1 or 0.

INTRODUCTION

This paper considers a financial scheduling problem introduced by Russell [2] in his paper, "Cash Flows in Networks." Given a project graph (of CPM or Pert type), the problem is one of scheduling events to maximize the present value of the outlays or receipts that occur at each event. Russell proposed a solution procedure and presented interesting interpretations of the problem, duality relations, and the solution procedure itself.

The principle results of this paper show that the scheduling problem (a nonlinear, nonconcave program) is equivalent to a linear program. This fact is then used to demonstrate that the optimal solution of the scheduling problem corresponds to a tree in the project graph, i.e., an extreme point in the set of feasible schedules. It follows that every local maximum is indeed a global maximum; thus the Kuhn-Tucker conditions are necessary and sufficient for optimality. This in-depth examination of the problem's special structure leads to an efficient solution procedure. The algorithm is related to Markowitz's special procedure for the weighted distribution problem [1, pp. 424–431]. We introduce a project deadline parameter into Russell's model. Our first algorithm solves the problem for any given project deadline. A second procedure finds the optimal solution for all possible deadlines. This yields a curve explicitly showing the tradeoff in project duration and present value.

There are four sections and an appendix. The appendix contains an example which illustrates most of the points in the paper. The first two sections discuss the problem and the equivalent linear program. The last two sections are given over to solution procedures.

† This work was supported, in part, by the National Science Foundation, Grant GS–3143.

THE PROBLEM

This section describes the payment scheduling problem while introducing notation and definitions. The concepts are illustrated by an example in the appendix. Readers who are unfamiliar with project graphs should consult Weist and Levy's excellent book [3].

$(V, \mathscr{A})$ is a directed graph. The nodes, $V=\{1, 2, . . ., N\}$, correspond to events, and the directed arcs, $\mathscr{A}$, correspond, with one exception, to activities.

Events 1 and N are special. Event 1 is the initiation of the project, and event N is the completion of the project. Throughout, collections of nodes will be designated by capital Roman letters, collections of arcs by capital script letters.

A real number, D_{ij}, is associated with each activity $(i, j) \in \mathscr{A}$; D_{ij} is the duration of activity (i, j). With one exception the D_{ij} are nonnegative. The special arc, $N1$, links the completion event N and initiation event 1. D_{N1} is a negative number: $-D_{N1}=\delta$ is an upper bound on the duration of the entire project.

Let D represent the vector with elements D_{ij} for $(i, j) \in \mathscr{A}$. The triple $(V, \mathscr{A}, D)$ forms a *project graph* if: (i) each event i is on a directed path from event 1 to event N; (ii) each directed cycle contains arc $(N, 1)$ and, the sum of the durations around any directed cycle is nonpositive.

It is useful to study the set of events immediately preceding and following each event. Let P_i be the set of events immediately preceding event i and F_i the events immediately following event i; j is in P_i if and only if $(j, i) \in \mathscr{A}$; j is in F_i if and only if $(i, j) \in \mathscr{A}$. When $(V, \mathscr{A}, D)$ is a project graph, the sets P_i and F_i are nonempty, $P_1=\{N\}$, $F_N=\{1\}$, and we can assume that $j \epsilon P_i (i \neq 1)$ implies $j< i$.

Variables T_i for $i=1, 2, . . ., N$ are associated with each event. T_i is called the epoch or occurrence time of event i. For $i \neq N$, T_i is the instant at which activities (i, j) for $j \in F_i$ are initiated. Activity (i, j) is then completed at epoch T_i+D_{ij}. We impose the condition that activities starting at event j cannot begin until all activities initiated at events $i \in P_j$ have been completed. Algebraically, this means

$$T_j - T_i \geqq D_{ij} \qquad \text{for } j=2, 3, . . ., N \quad i \in P_j.$$

The duration of the project is $T_N - T_1$. The duration is limited by

$$T_1 - T_N \geqq D_{N1}.$$

We always assume $T_1=0$, and call the $N-1$ vector $T=(T_2, T_3, . . ., T_N)$ a *schedule*. Note the last constraint becomes $T_N \leqq \delta$.

Let E be the node arc incidence matrix for $(V, \mathscr{A})$ with the (redundant) row corresponding to event 1 deleted. The schedule constraints are

$$TE \geqq D.$$

If the D_{ij} are viewed as distances, the length of the longest path from event 1 to event N gives the briefest possible duration of the project. Let δ_0 be that duration. The fact that $(V, \mathscr{A}, D)$ is a project graph indicates that $\delta_0 \leqq \delta$. The early start time ([3], pg. 27) for event i is the earliest possible epoch at which event i could occur; it is equal to the length of the longest path from node 1 to node i. The late

start time ([3], page 29) is the latest possible epoch of event i that is consistent with completion of the project in minimum time ($T_N = \delta_0$).

At epoch T_i a sum of money q_i changes hands; q_i is positive for receipts, negative for payments. The value of money is discounted at rate α. Thus the present value of the transaction at epoch i is $q_i \exp(-\alpha T_i)$. We assume $q_1 = 0$. The present value of any schedule T is given by

$$f(T) = \sum_{i=1}^{N} q_i \exp(-\alpha T_i).$$

The payment scheduling problem asks us to

$$\text{(I)} \qquad \text{Maximize} \quad f(T)$$

$$\text{Subject to} \quad TE \geqq D.$$

Notice three things about problem (I): (i) it has a feasible solution if and only if $\delta_0 \leqq \delta$; (ii) if $\delta = +\infty$ (no time limit on the project), then the maximum may not be achieved; e.g., if $q_N = -1$; and (iii) when the q_i differ in sign the objective function can fail to be concave (or pseudoconcave); local maxima might exist.

THE TRANSFORMED PROBLEM

In this section, it is demonstrated that problem (I) is equivalent to a linear program.

The equivalent problem is

$$\text{(II)} \qquad \text{Maximize} \quad \sum_{i=1}^{N} y_i q_i$$

$$\text{Subject to} \quad y_j K_{1j} \leqq 1 \qquad \text{for } j \in F_1$$

$$y_j K_{ij} - y_i \leqq\neq 0 \qquad \text{for } i = 2, \ldots, N;\ j \in F_i$$

$$-y_N \leqq -K_{N1}.$$

The constants K_{ij}, $(i, j) \in \mathscr{A}$, are defined by $K_{ij} = \exp(\alpha D_{ij})$; thus $K_{ij} \geqq 1$ if $(i, j) \neq (N, 1)$ and $0 < K_{N1} < 1$. The constraints of (I) are $T_j - T_i \geqq D_{ij}$ for $(i, j) \in \mathscr{A}$, where $T_1 = 0$. The constraints of (II) are then $y_j K_{ij} - y_i \leqq 0$ for $(i, j) \in \mathscr{A}$, where y_1 is equal to 1.

Let $y = (y_2, y_3, \ldots, y_N)$ be an $N-1$ vector. We define a one-to-one correspondence between the nonnegative vectors $T \geqq 0$ and the positive vectors y, where $y_i \leqq 1$ for all i:

$$y_i = \exp(-\alpha T_i)$$

or

$$T_i = \ln(y_i) / -\alpha$$

Let $y(T)$ be the y determined by T.

THEOREM 1: *(Equivalence of (I) and (II)).*

(a) T *is a feasible solution of (I) if and only if* y(T) *is a feasible solution of (II).*

(b) T *is optimal if and only if* y(T) *is optimal.*

PROOF: Note that (b) follows easily from (a) since $f(T)=y(T)q$. Let $y=y(T)$. T satisfies the constraints of (I) if and only if

$$\begin{aligned}
&\ln(y_j)/-\alpha &&\geqq D_{1j} &&\text{for } j\epsilon F_1\\
&[\ln(y_j)/-\alpha]-[\ln(y_i)/-\alpha] &&\geqq D_{ij} &&\text{for } i=2,\ldots,N\\
& && && j\epsilon F_i\\
&+[\ln(y_N)/-\alpha]\leqq\delta.
\end{aligned}$$

The constraints can be rearranged to say

$$\begin{aligned}
&\ln y_j &&\leqq -\alpha D_{1j}=\ln(K_{1j}^{-1}) &&\text{for } j\epsilon F_1\\
&\ln(y_j/y_i)=\ln y_j-\ln y_i &&\leqq -\alpha D_{ij}=\ln(K_{ij}^{-1}) &&\text{for } i=2,\ldots,N\\
& && && j\epsilon F_i\\
&\ln y_N\geqq-\alpha\delta &&=\ln(K_{N1}).
\end{aligned}$$

Recall $K_{ij}=\exp(\alpha D_{ij})$ so $\ln K_{ij}=\alpha D_{ij}$; $\ln(K_{ij}^{-1})=-\alpha D_{ij}$.

The natural logarithm is a strictly monotone transformation. If u_1 and u_2 are positive numbers, then $\ln u_1<\ln u_2$ if and only if $u_1<u_2$. This implies our constraints are equivalent to

$$\begin{aligned}
&y_j &&\leqq K_{1j}^{-1} &&\text{for } j\epsilon F_1\\
&y_j/y_i &&\leqq K_{ij}^{-1} &&\text{for } i=2,\ldots,N\\
& && && j\epsilon F_i\\
&y_N &&\geqq K_{N1}.
\end{aligned}$$

We see that these are precisely the constraints of (II): multiply constraint j in the first set by K_{1j}, multiply constraint (i,j) in the second set by y_iK_{ij}, and multiply the last constraint by -1. Q.E.D.

Let G be a matrix obtained from E as follows: if an element of E is 0 or -1, the corresponding element of G is 0 or -1; if an element of E is $+1$ (in column (ij) and row j), then the corresponding element of G is $K_{ij}=\exp(\alpha D_{ij})$.

Let c be a vector defined on $\mathscr{A}$, $c_{1j}=1$ for $j\epsilon F_1$, $c_{N1}=-K_{N1}$, otherwise, $c_{ij}=0$. Problems (II) become

(II) $$\text{Maximize} \quad yq$$

$$\text{Subject to} \quad yG \leqq c.$$

THEOREM (2): If $\delta o \leqq \delta < +\infty$; T is an extreme point of $\{T \mid TE \geqq D\}$ if and only if $y(T)$ is an extreme point of $\{y \mid yG \leqq c\}$.

PROOF: Both E and G have rank $N-1$, thus y and T will be extreme points if and only if they satisfy at least $N-1$ linearly independent constraints with equality. Recall there is a one to one correspondence between constraints (columns of G and E) and arcs of $(V, \mathscr{A})$.

A tree in the project graph $(V, \mathscr{A})$ is a subgraph of $N-1$ arcs and no loops. Let $\mathscr{T}$ be the arcs in the tree and $\mathscr{T}' = \mathscr{A}/\mathscr{T}$ the arcs deleted from $\mathscr{A}$ to form the tree. A tree is feasible if we can associate a schedule T with the tree so that

$$T_j - T_i = D_{ij} \qquad \text{for } (i, j) \in \mathscr{T}$$

and

$$T_j - T_i \geqq D_{ij} \qquad \text{for } (i, j) \in \mathscr{T}',$$

where $T_1 = 0$. It is well known that the relations $T_1 = 0$ and $T_j - T_i = D_{ij}$ for $(i, j) \in \mathscr{T}$ uniquely determine the schedule T. Every extreme point of $\{T \mid TE \geqq D\}$ is equivalent to a feasible tree.

Let T be determined by the feasible tree $(V, \mathscr{T})$. Thus, according to Theorem 1, $y(T)$ is in the set $\{y \mid yG \leqq c\}$. We can show $y(T)$ is an extreme point by demonstrating that the $N-1$ constraints corresponding to arcs in $\mathscr{T}$ are linearly independent. Recall that $y(T)$ satisfies a constraint with equality if and only if T satisfies the corresponding constraint with equality. To show the constraints are linearly independent, we show that

$$\pi_j K_{ij} - \pi_i = 0 \qquad (i, j) \in \mathscr{T}$$

$$\pi_1 = 0$$

has only one solution $(\pi_2, \ldots, \pi_n) = 0$. Since $(V, \mathscr{T})$ is a tree, some arc is incident on node 1. For these arcs

$$\pi_j K_{1j} = \pi_1 = 0,$$

or

$$0 = \pi_1 K_{n1} = \pi_n.$$

Since $K_{ij} \neq 0$, $\pi_j = 0$ for all nodes in $\mathscr{T}$ incident on node 1. This argument can be continued by using the connectedness of $(V, \mathscr{T})$ to show $\pi_i = 0$ for all i. Therefore, $y(T)$ is an extreme point.

For the converse, suppose y is an extreme point. Since $\delta < \infty$, $K_{n1} > 0$, and $y_n \geqq K_{n1} > 0$. Each

event is on a directed path from 1 to N, so $1 \geqq y_j \geqq y_n > 0$ for all j. Let $\mathscr{S}$ be the arcs corresponding to the $N-1$ tight, linearly independent constraints and let $\tilde{G}$ be the corresponding $(N-1)$ x $(N-1)$ submatrix of G. If $(V, \mathscr{S})$ is not a tree, the subgraph $(V, \mathscr{S})$ is disconnected. If $(V, \mathscr{S})$ contains an isolated node, then $\tilde{G}$ has a zero row contradicting linear independence.

Consider a connected component that *does not* contain node 1. For arcs in this component $y_j K_{ij} = y_i$, because $c_{ij} = 0$ unless i or j equals 1. Since $\tilde{G}$ is nonsingular, the only solution for these relation is $y_i = 0$ for nodes in the connected component. This is impossible since $y > 0$. Therefore, $(V, \mathscr{S})$ is a tree. Q.E.D.

As a consequence of Theorem 2, we can restrict our search for optimal schedules to feasible trees in the project graph. The algorithm described below does just that, while following the logic of the simplex method applied to problem (II).

Given a feasible tree $(V, \mathscr{T})$ we know T and $y(T)$. The optimality of $(V, \mathscr{T})$ is checked, using the standard complementary slackness results. Let $\tilde{G}$ be the submatrix of G corresponding to arcs in $\mathscr{T}$. If the $N-1$ vector x solving $\tilde{G}\, x = q$ is nonnegative, then the current solution is optimal. If some element of x is strictly negative, then the corresponding arc can be deleted from the tree and a better tree can be obtained.

Recall that $\mathscr{T}' = \mathscr{A} / \mathscr{T}$. We find x by setting $x_{ij} = 0$ for $(i, j) \in \mathscr{T}'$ and then solving

(1)
$$\sum_{j \in P_i} K_{ji} x_{ji} - \sum_{j \in F_i} x_{ij} = q_i$$

for $i = 2, 3, \ldots, N$.

As an alternative to solving (1), we can, given T and thus $y(T)$, set $b_i = q_i y_i$ and solve

(2)
$$\sum_{j \in P_i} w_{ji} - \sum_{j \in F_i} w_{ij} = q_i y_i = b_i$$

for $i = 2, 3, \ldots, N$.

For any feasible tree set $x_{ij} = w_{ij} = 0$ for $(i, j) \in \mathscr{T}'$. Let y be the extreme solution of II corresponding to $\mathscr{T}$. For $(i, j) \in \mathscr{T}$ set $x_{ij} y_i = w_{ij}$.

THEOREM (3): x_{ij} solves (1) if and only if w_{ij} solves (2).

PROOF: Substitute $x_{ij} y_i$ for w_{ij} in (2), and divide each equation by $y_i > 0$.

We obtain

$$\sum_{j \in P_i} (y_j / y_i)\, x_{ji} - \sum_{j \in F_i} x_{ij} = q_i$$

for $i = 2, \ldots, N$.

If $x_{ji} > 0$, then $(j, i) \in \mathscr{T}$ and $y_i K_{ji} = y_i$, so $y_j / y_i = K_{ji}$. Therefore, x_{ij} solves (1). Note the steps are reversible if x_{ij} solves (1) then $K_{ji} x_{ji} = (y_j / y_i) x_{ji} = w_{ji} / y_i$. Q.E.D.

Since $w_{ij} \geqq 0$ if $x_{ij} \geqq 0$, optimality can be checked by solving (2). This saves work since (2) is a triangular system with only 0, and ± 1 for coefficients. If we find some $w_{lk} < 0$ for $(l, k) \in \mathscr{T}$ then the current tree is not optimal, removing arc (l, k) from $\mathscr{T}$ will disconnect the graph $(V, \mathscr{T})$. Suppose

$(U, \mathscr{S})$ and $(W, \mathscr{R})$ are the disconnected subgraphs and node 1 is in $(U, \mathscr{S})$. Two cases can be considered.

CASE 1: Node l is in U node $K \in W$. In the current schedule, we have $T_k - T_l = D_{lk}$. If the new T^* is going to be feasible, we must have $T^*_k - T^*_l \geqq D_{lk}$. One arc from $\mathscr{T}'$ will be added to $(U, \mathscr{S})$ and $(W, \mathscr{R})$ to form the new tree. Obviously, this new arc will connect the disjoint subgraphs. All the arcs in $\mathscr{S}$ will be in the new tree. From this we can conclude that $T^*_i = T_i$ for all $i \in U$. This is true since all T_i for $i \in U$ are determined by $T_j - T_i = D_{ij}$ for $(i, j) \in \mathscr{S}$ and $T_1 = 0$.

In particular, T_l will not change; $T^*_l = T_l$. This means $T^*_k \geqq T_k$ if the new tree is going to be feasible. If T_K increases, then to preserve the equation, $T_j - T_i = D_{ij}$ for $(i, j) \in \mathscr{R}$, we must have each T_i increasing (for $i \in W$) by the same amount as T_K. Thus the question is how much can we increase the T_i for $i \in W$ and still remain feasible. The only arcs we have to worry about are those running from W to U. The limit of the increase is given by

$$\theta = \text{Min } [T_j - T_i - D_{ij} | i \in W, j \in U, (i, j) \in \mathscr{T}'].$$

Suppose arc (m, n) achieves the minimum, then $\theta = T_m - T_n - D_{mn} \geqq 0$. Arc (m, n) is added to $\mathscr{S}$ and $\mathscr{R}$ to form a new tree.

The new values of T and y are readily calculated.

CASE 2: $l \in W$, $k \in U$. The same reasoning applies. T_k must remain constant. All the T_i for $i \in W$ must decrease by a constant amount. Only arcs $(i, j) \in \mathscr{T}$ running from $(U, \mathscr{S})$ to $(W, \mathscr{R})$ need to be considered. The decrease is given by

$$\theta = \text{Max } [T_i + D_{ij} - T_j | i \in U, j \in W, (i, j) \in \mathscr{T}'].$$

If arc (m, n) achieves the maximum, then the new tree is formed by adding (m, n) to $\mathscr{S}$ and $\mathscr{R}$. Note that $\theta = T_m + D_{mn} - T_n$ is nonpositive. In either case for $i \in W$, the new epoch is $T_i + \theta$. Therefore, the new b vector for $i \in W$ is given by

$$q_i \exp [-\alpha (T_i + \theta)] = q_i \exp [-\alpha T_i] \exp [-\alpha \theta] = b_i \Delta,$$

where

$$\Delta = \exp [-\alpha \theta].$$

The increase in present value is $(\Delta - 1) \sum_{i \in W} b_i$.

PROPOSITION: The increase in present value is given by

(i) $(\Delta - 1)\, w_{lk}$ in case 1, and

(ii) $-(\Delta - 1)\, w_{lk}$ in case 2.

PROOF: We must show $\sum_{i \in W} b_i$ equals w_{lk} in case 1 and $-w_{lk}$ in case 2. To determine this sum, the relation in (2) for $i \in W$. If an arc in $\mathscr{T}$ runs from U to U, it is not counted. If it runs from W to W, it is counted once in a positive sense and once in a negative sense. Arc (l, k) is the only arc in $\mathscr{T}$ connecting U and W. If $l \in U$ and $k \in W$ (case 1), then at node $k \in W$, we obtain a contribution w_{lk}. Therefore,

$$\sum_{i\epsilon W} b_i = \sum_{i\epsilon W} [\sum_{i\epsilon P_i} w_{ji} - \sum_{j\epsilon F_i} w_{ij}] = \begin{Bmatrix} w_{lk} \text{ in case 1} \\ -w_{lk} \text{ in case 2} \end{Bmatrix}.$$

Note, if $\theta = 0$, that a positive increase in present value is made at each step. In case (1) $\theta > 0$, thus $\Delta < 1$, so $(\Delta - 1) < 0$, $w_{lk} < 0$ implies $(\Delta - 1)\ w_{lk} > 0$. In case 2 $\theta < 0$, $\Delta > 1$, so $-(\Delta - 1) < 0$, $w_{lk} < 0$ implies $-(\Delta - 1) w_{lk} > 0$. **Q.E.D.**

Note the economic interpretation. In case 1, we have isolated a set of nodes (W, not containing node 1) with a net loss of present value. The algorithm directs us to delay that loss as much as possible. In case 2, we find a set of nodes (W) with a net gain in present value. The algorithm directs us to advance the scheduling of events in W.

Let $PV(D)$ be the optimal value of problem (I) as a function of the duration vector D. As Russell has indicated, the w_{ij} can be interpreted as

$$-\frac{\partial PV(D)}{\partial D_{ij}} = w_{ij}/\alpha.$$

Thus w_{ij}/α is the marginal present value return of shortening the duration of activity (i, j). If $D^*_{ij} = D_{ij} - \Delta$, then $PV(D^*) \cong PV(D) + (\Delta, \alpha) w_{ij}$. Since $x_{ij} y_i = w_{ij}$, this indicates

$$x_{ij}/\alpha = -\frac{\partial PV(D)}{\partial D_{ij}} \cdot \exp(+\alpha T_i).$$

The x_{ij}/α measures the marginal return at time T_i that is obtained by shortening the duration of activity (i, j).

THE FIXED DEADLINE ALGORITHM

PHASE 1: Find the early tree ([3], pg 27) $(V, \mathscr{T})$ and the early start times T. Set $T_N = \delta_0$; if $\delta_0 > \delta$ no schedule can meet the deadline. Otherwise calculate $b_i = q_i \exp(-\alpha T_i)$ for $i = 1, 2, \ldots N$; set $PV = \sum_{i=2}^{N} b_i$.

PHASE 2:

STEP 1. *Finding an arc to delete from the feasible tree.* Set $w_{ij} = 0$ for $(i, j) \epsilon \mathscr{T}'$, then solve

$$\sum_{j\epsilon P_i} w_{ij} - \sum_{j\epsilon F_i} w_{ij} = b_i \qquad \text{for } i = 2, 3, \ldots, N.$$

If all $w_{ij} \geqq 0$, the current solution is optimal. If some $w_{lk} < 0$, go to step 2.

STEP 2. *Finding a new arc.* Deleting arc (l, k) from $(V, \mathscr{T})$ forms subgraphs $(U, \mathscr{S})$ and $(W, \mathscr{R})$ where node 1 is in U. If $l \epsilon U$ go to step 3; if $l \epsilon W$ go to step 4.

STEP 3. *Delay the events in* W. Compute

$$\theta = \text{Min}[T_j - T_i - D_{ij} \mid i \epsilon W, j \epsilon U, (i, j) \epsilon \mathscr{T}'].$$

Suppose arc (m, n) achieves the minimum. Go to step 5.

STEP 4. *Advance the events in* W. Compute

$$\theta = \mathrm{Max}[T_i + D_{ij} - T_j \mid i \in U, j \in W, (i, j) \in \mathcal{T}'].$$

Suppose arc (m, n) achieves the maximum. Go to step 5.

STEP 5. *Change of schedule.* Let $\Delta = \exp(-\alpha\theta)$. For $i \in W$,

$$T_i \leftarrow T_i + \theta;$$

$$b_i \leftarrow b_i \Delta;$$

Delete (l, k) from $\mathcal{T}$ and add (m, n). Go to step 1.

We can note that:

(i). If $\delta = \delta_0$, the arc $(N, 1)$ is likely to be in the optimal tree. In this case, Phase 1 could be modified as follows: find the late tree, delete the arc on the critical path joining node 1, add arc $(N, 1)$, set T equal to the late start times.

(ii). Sensitivity of the optimal schedule to changes in α can be determined from (2). Keep T fixed set $w_{ij} = 0$ for $(i, j) \in \mathcal{T}'$, then determine the range of α for which (2) has a nonnegative solution. For example if α is increased until some w_{ij} becomes 0, then the algorithm described above can be used to find the new optimal tree. Unfortunately this sensitivity procedure involves some difficult calculations.

THE PARAMETRIC ALGORITHM

It may be desirable to solve problem (I) parametrically in the deadline constraint to measure the exact trade-off between present value and completion time. This section presents such an algorithm.

We assume the procedure described in the last section has been used to find an optimal schedule for $\delta = \delta_0$. Then, δ is increased to $+\infty$. One of two things will happen; either (i) the optimal solution will become insensitive to further increases in the deadline, or (ii) we will discover that the deadline constraint will be binding for any limit. This implies it is not optimal to complete the project if no deadline is imposed.

We find that the optimal value $PV(\delta)$ is a concave increasing function of δ made up of exponential segments. If a tree is optimal in the region $\delta_h \leqq \delta \leqq \delta_{h+1}$ and W_{N1} is in the solution related to this tree, then

$$PV(\delta) = [(1 - \exp(-\alpha[\delta - \delta_h]))]\alpha \cdot w_{N1} + PV(\delta_h)$$

for $\delta_h \leqq \delta \leqq \delta_{h+1}$.

The parametric procedure follows the logic of the simplex method while exploiting the special structure of our problem. We assume Phase 1 and Phase 2 have been completed.

PHASE 3:

STEP 0. *Start.* From Phase 1 and 2 we know δ_0 and the optimal tree $(V, \mathcal{T})$ when $\delta = \delta_0$. Let PV, T, and w be the solution obtained from Phase 2. Set $h = 0$, $PV_0 = PV$.

STEP 1: *Optimality Check, Find arc to join tree.* If $w_{N1} = 0$, the current schedule T is optimal for all $\delta \geqq \delta_h$. Otherwise, delete arc $(N, 1)$ from $(V, \mathcal{T})$ and let $(U, \mathcal{S})$ and $(W, \mathcal{R})$ be the disconnected subgraphs; $1 \in U$, $N \in W$. If no arc, aside from $(N, 1)$, goes from W to U, then the current tree is optimal

for all $\delta \geqq \delta_h$. Let $\theta = \delta - \delta_h$. The optimal schedule is

$$T_i(\theta) = \begin{Bmatrix} T_i, & \text{if } i \in U \\ T_i + \theta, & \text{if } i \in W \end{Bmatrix}.$$

If some arc goes from W to U, compute

$$\theta = \mathrm{Max}[T_j - T_i - D_{ij} \mid i \in W, j \in U, (i, j) \in \mathscr{T}'].$$

Suppose $\theta = T_n - T_m - D_{mn} \geqq 0$.

STEP 2. *New solution.* Set

$$\Delta \leftarrow \exp(-\alpha\theta)$$

$$\delta_{h+1} \leftarrow \delta_h + \theta$$

$$PV_{h+1} \leftarrow PV_h + (1 - \Delta) w_{N1}$$

$$h \leftarrow h + 1.$$

For $i \in W$, set

$$T_i \leftarrow T_i + \theta$$

$$W_{ij} \leftarrow w_{ij}\Delta.$$

STEP 3. *Finding an arc to drop.* Adding (m, n) to $(V, \mathscr{T})$ forms a circuit $\mathscr{C}$. Note that (m, n) and $(N, 1)$ are in $\mathscr{C}$ and that the arc oriented in opposite fashions. Let $\mathscr{C}'$ be all arcs in $\mathscr{C}$ oriented as (m, n) and $\mathscr{C}^2$ the arcs oriented as $(N, 1)$. Find

$$\theta = \mathrm{Min}\,\{w_{ij} \mid (i, j) \in \mathscr{C}^2\}.$$

Suppose $\theta = w_{lk} \geqq 0$.
For $(i, j) \in \mathscr{C}^1$

$$w_{ij} \leftarrow w_{ij} + \theta;$$

For $(i, j) \in \mathscr{C}^2$

$$w_{ij} \leftarrow w_{ij} - \epsilon,$$

go to step 1.

APPENDIX

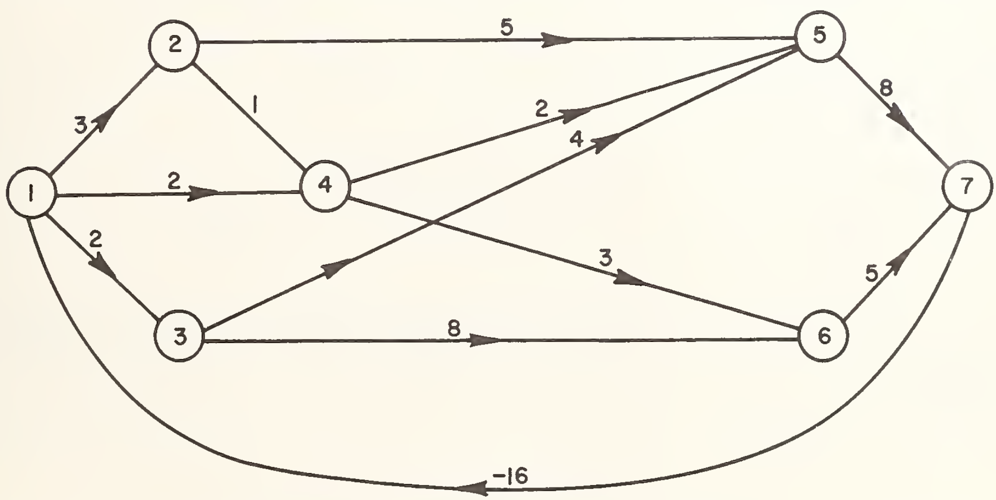
AN EXAMPLE

This section gives an example of payment scheduling problems. There are 7 events, 11 activities, and a deadline on the project. Thus the project graph has 7 nodes and 12 arcs.

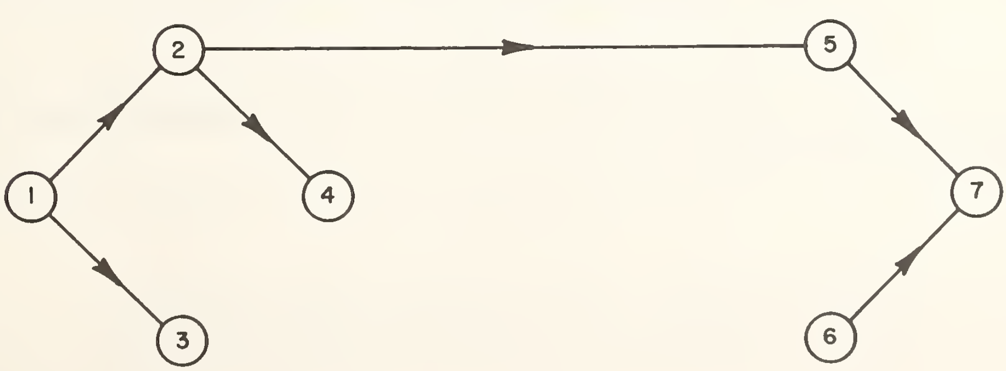
Data for the problem are summarized in the table and figure following. We are using the same problem as Russell [2, p. 370]. The discount factor is $\alpha = 0.01$; time is measured in months. Russell assumes $q_1 = -100$. We take $q_1 = 0$. Thus the value of our solution will be 100 greater than Russell's.

Event	P_i	F_i	q_i	Early starts
1	7	2, 3, 4	0	0
2	1	4,5	−200	3
3	1	5, 6	−200	2
4	1, 2	5, 6	100	2
5	2, 3, 4	7	400	8
6	3, 4	7	−200	10
7	5, 6	1	300	16

The network is drawn below; the numbers on the arcs are the D_{ij}. Note the deadline is 16.

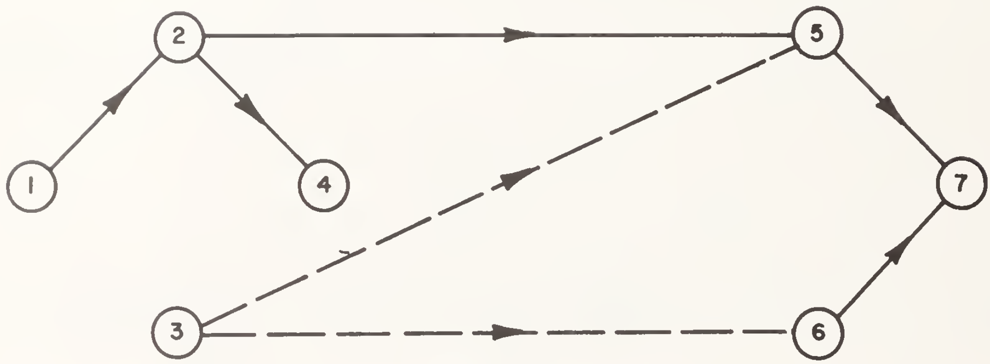


We show one iteration of the fixed deadline algorithm. A feasible tree and the associated values of T and b are shown below. The present value of this solution is 151.69.



	2	3	4	5	6	7
T	3	2	4	8	11	16
b	−194.09	196.04	96.08	369.24	−179.16	255.66

On solving (2) we find that $w_{13} = -196.04$. Arc (1, 3) is detached from the tree. The possible incoming arcs are dotted.



According to Step 3,

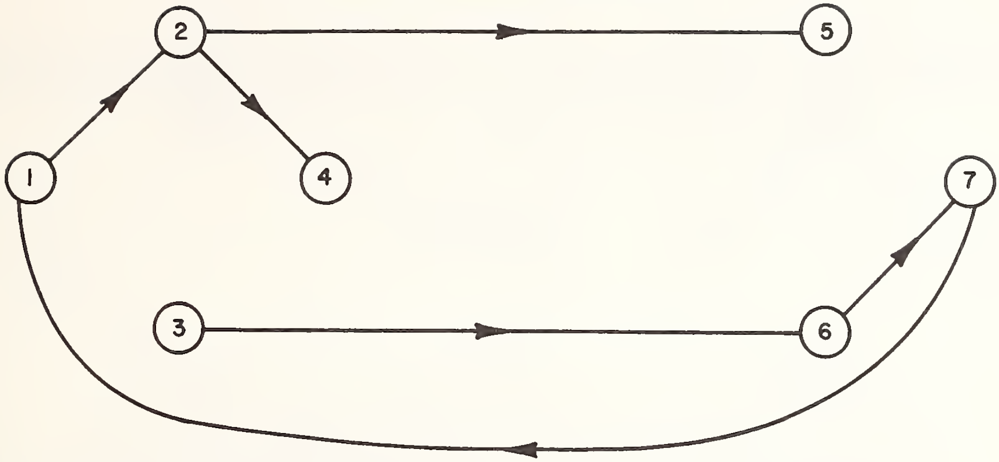
$$\theta = \text{Min}\begin{bmatrix} T_5 - T_3 - D_{35} = 2 \\ T_6 - T_3 - D_{63} = 1 \end{bmatrix} = 1.$$

Arc (3, 6) joins *th* tree. The new values of b and T are tabulated below. The present value of the new schedule is 153.64.

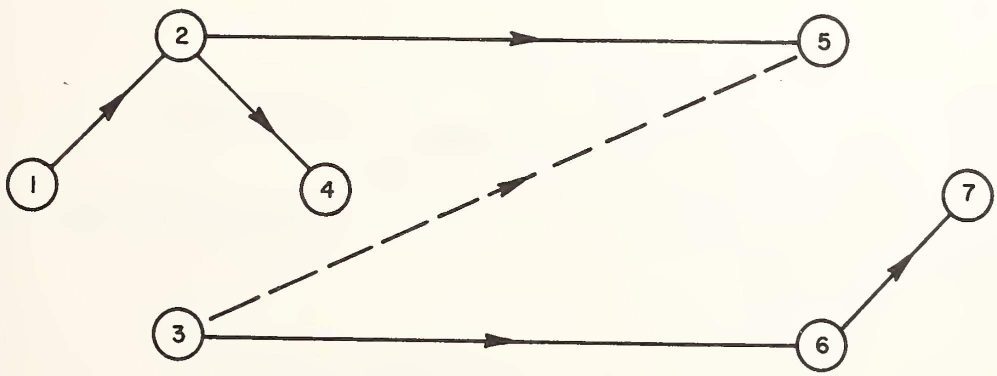
	2	3	4	5	6	7
T	3	3	4	8	11	16
b	−194.09	−194.09	96.08	369.24	−179.16	255.66

These values of T and b are optimal although we must go through one degenerate step to get the optimal values of w and the final tree.

	(1,2)	(2,4)	(2,5)	(3,6)	(6,7)	(7,1)
w	271.23	96.08	369.24	194.09	373.25	117.59



Now, we apply the parametric algorithm to our problem: $\delta_0 = 16$ and $PV_0 = 153.64$. Since $w_{71} = 117.59 > 0$, we look at the subgraphs formed by deleting (7,1).



The only arc, except 71, from $(W, \mathscr{R})$ to $(U, \mathscr{S})$ is (3,5). Thus

$$\theta = T_5 - T_3 - D_{35} = 1.$$

In Step 2, we find the new values $\Delta = 0.99$, $\delta_1 = 17$, and $PV_1 = 154.80$. The new values of T and w are

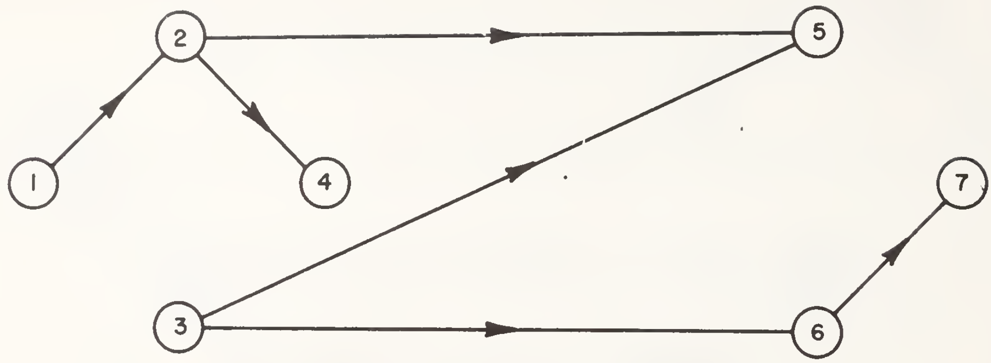
	2	3	4	5	6	7
T	3	4	4	8	12	17

	(1,2)	(2,4)	(2,5)	(3,6)	(6,7)	(7,1)
w	271.23	96.08	369.24	192.16	369.54	116.43

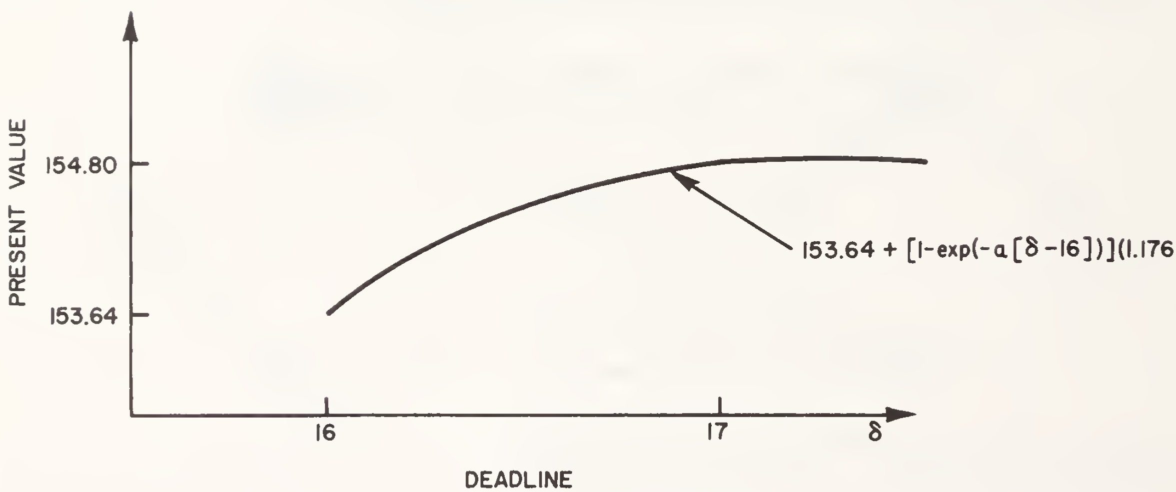
Notice that $PV_1 = w_{12} - w_{71}$. Adding arc (3,5) creates a circuit with $C^1 = \{(3,5)\}$; $C^2 = \{(7,1), (1,2), (2,5), (3,6), (6,7)\}$. Thus $\theta = w_{71} = 116.43$ and the new values of w are

	(1,2)	(2,4)	(2,5)	(3,6)	(6,7)	(3,5)
w	154.79	96.08	252.81	75.73	253.11	116.43

The tree is



Since $w_{71}=0$, this solution is optimal for all values of $\delta \geqq 17$. The plot of present value against deadline is



ACKNOWLEDGMENT

I express my appreciation to the referee for his thorough report.

REFERENCES

[1] Dantzig, G. B., *Linear Programming and Extensions* (Princeton University Press, New Jersey, 1963).
[2] Russell, A. H., "Cash Flows in Networks," Management Science, *16*, 357–73 (Jan. 1970).
[3] Weist, J. D. and F. K. Levy, *A Management Guide to PERT/CPM* (Prentice-Hall, 1969).

SEQUENTIAL BID SELECTION BY STOCHASTIC APPROXIMATION

Robert A. Agnew

Montgomery Ward
Chicago, Illinois

ABSTRACT

Suppose that a contractor is faced with a sequence of "minimum bid wins contract" competitions. Assuming that a contractor knows his cost to fulfill the contract at each competition and that competitors are merely informed whether or not they have won, bids may be selected sequentially via a tailored stochastic approximation procedure. The efficacy of this approach in certain bidding environments is investigated.

1. INTRODUCTION

Consider the sequential decision problem of a contractor facing an interminable sequence of "minimum bid wins contract" (sealed-bid) competitions for similar contracts. If a random bidding environment is assumed where the competitions are independent and the probability of winning any competition is merely a function of percentage markup over cost, the contractor will naturally attempt to approximate a suitable percentage markup via some sequential approximation procedure that incorporates the data available from past competitions. The form of the data available depends on the information divulged to the competitors after each competition. Under minimal disclosure, a competitor is merely informed whether or not he has won.

In this paper, we suggest a Kiefer-Wolfowitz type stochastic approximation procedure for the purpose of converging on that percentage markup which maximizes expected percentage profit. The procedure assumes only minimal disclosure, and it converges under relatively weak assumptions regarding the p function, i.e., the probability of winning as a function of percentage markup. In the next section, we delineate those assumptions. In Section 3, the approximation procedure is presented and its convergence properties are discussed. Section 4 and the Appendix contain the results of a simulated numerical example designed to provide additional insight into the behavior of our approximation procedure. Section 5 discusses the addition of a penalty associated with losing a competition and the behavior of our procedure in a particular deterministic bidding environment. For general accounts of stochastic approximation, refer to [17], [22], and [23]. We shall make use of results in [2] and [3]. Generalizations of the Kiefer-Wolfowitz procedure are discussed in [5] and [6].

The quantitative competitive bidding literature contains papers developing the entire spectrum of mathematical models from purely statistical to purely game theoretic. We have included a variety of these papers in our references; however, our approach is purely statistical and is properly considered as a descendant of the original Friedman model [8]. Unlike previously suggested statistical procedures, however, it is nonparametric, i.e., it is not necessary to specify a functional form for p. The calculations involved are trivial; they can be performed on any modern desk calculator, or for that matter, on a slide rule.

2. PRELIMINARIES

For $x \geqslant -100$, let $p(x)$ be the probability of winning any competition when the percentage markup over cost is x, i.e., given cost $C > 0$, the bid is $(1+0.01x)C$, and p is assumed to be independent of C. (By cost here, we mean the direct projected expenditures involved in fulfilling the terms of the contract.) This setup is equivalent to that of the Friedman model.

We assume that p is nonincreasing on $(-100, \infty)$ and that $b = \sup\{x : p(x) > 0\} \epsilon (0, \infty)$. It is clear that there exists, in any realistic situation, a finite least upper bound on percentage markups with any chance of winning. We do not assume that its value is known, but we do assume that $0 < b_1 \leqslant b \leqslant b_2 < \infty$ with b_1, b_2 known. In other words, we assume that the contractor is able to specify a compact interval in $(0, \infty)$ which covers b.

We shall assume that the expected percentage profit function $f(x) = xp(x)$ is strictly unimodal with unique maximum point $\theta \epsilon (0, b)$. Moreover, we assume that f is strictly increasing on $(0, \theta)$, with positive lower derivative bounded away from zero on $(0, \theta - \delta)$ for arbitrarily small $\delta > 0$, and that f is strictly decreasing on (θ, b), with negative upper derivative bounded away from zero on $(\theta + \delta, b)$ for arbitrarily small $\delta > 0$. The conditions on f are required in order to disallow such phenomena as zero-slope inflection points as well as multiple local maxima. The strict unimodality of f implies, in particular, that p must be continuous on $(0, \theta)$, left continuous at θ, and strictly decreasing on (θ, b).

3. THE APPROXIMATION PROCEDURE

Choose positive numbers X_1, a, c, and α such that $X_1 < b_1$, $c < \min(X_1, b_1 - X_1)$, $a \leqslant 2c^2/b_2$, and $3/4 < \alpha < 1$. Define positive sequences $\{a_n\}$ and $\{c_n\}$ by $a_n = an^{-\alpha}$ and $c_n = cn^{-1/4}$. With U_n^+ and U_n^- depending only on X_n, let

(1) $$P\{U_n^+ = 1 \mid X_n\} = 1 - P\{U_n^+ = 0 \mid X_n\} = p(X_n + c_n)\,,$$

(2) $$V_n^+ = (X_n + c_n)U_n^+\,,$$

(3) $$P\{U_n^- = 1 \mid X_n\} = 1 - P\{U_n^- = 0 \mid X_n\} = p(X_n - c_n)\,,$$

(4) $$V_n^- = (X_n - c_n)U_n^-\,,$$

(5) $$Y_n = (2c_n)^{-1}(V_n^+ - V_n^-)\,,$$

(6) $$X_{n+1} = X_n + a_n Y_n\,,$$

(7) $$W_n = (2n)^{-1} \sum_{i=1}^{n} (U_i^+ + U_i^-)\,, \text{ and}$$

(8) $$Z_n = (2n)^{-1} \sum_{i=1}^{n} (V_i^+ + V_i^-)\,.$$

In the above setup, X_n is the estimate or approximation of θ just prior to the nth stage, and (6) is the basic recursion relation. The nth stage consists of two consecutive competitions at percentage

markups X_n+c_n and X_n-c_n yielding percentage profits V_n^+ and V_n^-, respectively; U_n^+ and U_n^- are just the zero/one (lose/win) indicator random variables associated with the nth stage competitions. Y_n is the estimate of the derivative or gradient of f at X_n. (f need not be differentiable; note, however, that $E(Y_n|X_n)=(2c_n)^{-1}(f(X_n+c_n)-f(X_n-c_n))$.) The estimate moves in the estimated gradient direction with the step size and the spread between percentage markups at a stage decreasing at predetermined rates as n increases. W_n is the proportion of competitions won and Z_n is the average percentage profit over the first n stages.

Our conditions on X_1, a, and c insure that $c_n < X_n < b$ for all n. Under the assumptions of Section 2, $X_n \to \theta$ as $n \to \infty$ with probability one and in the mean square, i.e., $P\{\lim X_n=\theta\}=1$ and $\lim E[(X_n-\theta)^2]=0$ (cf. [3]). Moreover, if p is continuous at θ, then $W_n \to p(\theta)$ and $Z_n \to f(\theta)$ with probability one. Some additional assumptions yield an asymptotic distribution for X_n. If $p(\theta) < 1$, p'' exists and is continuous in a neighborhood of θ, and $f''(\theta)<0$, then $(n^{\alpha-1/2})^{1/2}(X_n-\theta)$ is asymptotically normal with mean zero and variance

$$\sigma^2=(a/4c^2)(f(\theta)(\theta-f(\theta))/|f''(\theta)|). \tag{9}$$

In other words, X_n is approximately normally distributed with mean θ and variance $\sigma^2/n^{\alpha-1/2}$ for n sufficiently large. (This result follows from a theorem in [2] since $\lim\mathrm{var}(V_n^+|X_n)=\lim\mathrm{var}(V_n^-|X_n)=\theta^2p(\theta)(1-p(\theta))=f(\theta)(\theta-f(\theta))$. The general theorem in [6] yields the same result under a Lipschitz condition on p'' at θ.) Of course, σ^2 is unknown, but this result does indicate the speed of convergence and the dependence of asymptotic variance upon the various parameters. It is not surprising that σ^2 is inversely related to the degree of curvature or "peakedness" of f at θ.

The asymptotic performance of the procedure obviously improves as c is increased, as a is decreased, and as α is increased. (We have disallowed $\alpha=1$ because a special condition required in this case for asymptotic normality will not generally obtain under our assumptions.) Assuming that c is specified, the magnitude of step size is governed by a, and the rate of decrease in step size is governed by α. A procedure with relatively small a and α near one might be termed "conservative," and a procedure with relatively large a and α near three-quarters might be termed "aggressive." A conservative procedure is relatively good asymptotically, but it may not move much in the short run. An aggressive procedure is relatively bad asymptotically, but it may be preferable in the short run if the initial bias $|X_1-\theta|$ is large. A lot depends on the confidence one has in the initial estimate X_1. In the case of a seasoned competitor, this estimate may incorporate a good deal of experiential intuition and/or historical data. It may be a percentage markup that has yielded satisfactory results in the past. In such instances, conservatism may be appropriate. On the other hand, an aggressive stance may be appropriate for a relatively inexperienced competitor.

We note that there is no stopping mechanism built into the approximation procedure. It is assumed that the contractor continues the procedure indefinitely. Under our assumptions, he has no reason to do otherwise.

4. SIMULATED NUMERICAL EXAMPLE

In order to give a better idea of how the approximation procedure works, we have simulated three 25-stage runs under the assumption that $p(x)=0.8-0.04x$ for $0 \leq x \leq 20=b$, $b_1=15$, and $b_2=30$. Then $\theta=10$, $p(\theta)=0.4$, and $f(\theta)=4$. In each case, we have put $X_1=9$, $c=3$, $a=0.6$, and

$\alpha=0.76$. The random numbers employed were taken from [18] as follows: column 9, page 626 in Run 1; column 1, page 627 in Run 2; column 3, page 628 in Run 3. The results are rounded to three decimal places, although more digits were actually carried in the calculations. The runs are detailed in the Appendix. The net results are: $X_{26}=9.945$, $W_{25}=0.300$, $Z_{25}=2.877$ in Run 1; $X_{26}=10.055$, $W_{25}=0.480$, $Z_{25}=4.833$ in Run 2; $X_{26}=9.266$, $W_{25}=0.440$, $Z_{25}=3.796$ in Run 3.

These runs represent three possible (although not necessarily likely) realizations for the first 25 stages of an approximation procedure with specified parameter values in the assumed bidding environment. In each case, there is a net movement, however irregular, in the direction of θ. Actually, one should expect a lot of "wandering" in this example since the objective function is relatively "unpeaked."

5. CONCLUDING REMARKS

Suppose that the contractor wishes to associate a penalty with losing a competition. Presumably, such a penalty should vary with the size of the contract. Suppose that the penalty is a fixed percentage $\beta>0$ of the contract cost, i.e., the penalty cost is $0.01\beta C$ if the contract cost is C. Then, the contractor wishes to approximate the percentage markup θ maximizing

$$f(x)=xp(x)-\beta(1-p(x))=(x+\beta)p(x)-\beta, \tag{10}$$

for $x>-\beta$. The changes required in the approximation procedure of Section 3 are $c-\beta<X_1<b_1-c$, $V_n^+=(X_n+c_n+\beta)U_n^+-\beta$, and $V_n^-=(X_n-c_n+\beta)U_n^--\beta$.

In practice, a suitable objective value for β may be difficult to obtain. However, a positive penalty may be viewed merely as a subjective device to reflect the contractor's increasing aversion to losing contracts of increasing size. Increasing β has the conservative effect of decreasing the optimal percentage markup and of increasing the corresponding probability of winning.

In Section 2, we have made certain plausible assumptions regarding the p function which guarantee convergence of the approximation procedure. However, these assumptions are not necessary for the procedure to produce satisfactory results. Consider a deterministic bidding environment where $p(x)=1$ for $x\leq b\epsilon(0,\infty)$ and $p(x)=0$ for $x>b$. One can think of this situation as arising in two somewhat artificial ways. The contractor could have a single, fixed competitor who always bids a constant (unknown) percentage markup over the contractor's cost with ties decided in favor of the contractor. Alternatively, the contractor might be the sole bidder with a fixed (unknown) upper bound on acceptable percentage markup stipulated by the contractee. Using the procedure of Section 3 with b, b_1, and b_2 playing the same roles, it is not difficult to show that $X_n\rightarrow b$, $W_n\rightarrow 1$, and $Z_n\rightarrow b$ as $n\rightarrow\infty$. As n increases, the markups get generally closer to b, but fewer and fewer of them exceed b. Our previous remarks concerning the choice of a and α also apply here.

The practical man may question the utility of a mechanistic procedure which converges rather slowly. It is true that our procedure may be somewhat myopic in a bidding environment where full disclosure of bids and identities is the rule. Even in such an environment, however, our procedure might be useful as a final "honing" device. Regarding speed of convergence, we remark that statistical estimation procedures are usually not too fast. In terms of asymptotic variance, one can generally do no better than order n^{-1}. Finally, the contractor will have to adapt to the bidding environment in some fashion. Our procedure at least has the advantage of simplicity and of known asymptotic properties under relatively weak assumptions.

APPENDIX

TABLE 1. *Simulation Run 1*

n	X_n	U_n^+	V_n^+	U_n^-	V_n^-
1	9.000	0	0	0	0
2	9.000	0	0	0	0
3	9.000	0	0	0	0
4	9.000	0	0	0	0
5	9.000	1	11.006	1	6.994
6	9.177	0	0	0	0
7	9.177	1	11.021	1	7.332
8	9.313	0	0	1	7.530
9	9.053	1	10.785	0	0
10	9.404	1	11.091	0	0
11	9.747	0	0	0	0
12	9.747	0	0	1	8.135
13	9.518	0	0	0	0
14	9.518	0	0	0	0
15	9.518	0	0	0	0
16	9.518	0	0	0	0
17	9.518	0	0	0	0
18	9.518	1	10.974	0	0
19	9.769	1	11.206	1	8.332
20	9.833	0	0	0	0
21	9.833	1	11.235	1	8.432
22	9.893	0	0	0	0
23	9.893	0	0	0	0
24	9.893	0	0	0	0
25	9.893	1	11.234	1	8.551
26	9.945				

TABLE 2. *Simulation Run 2*

n	X_n	U_n^+	V_n^+	U_n^-	V_n^-
1	9.000	1	12.000	1	6.000
2	9.600	0	0	0	0
3	9.600	1	11.880	1	7.320
4	9.860	0	0	1	7.739
5	9.479	1	11.485	0	0
6	9.984	1	11.901	0	0
7	10.461	1	12.306	1	8.617
8	10.598	1	12.382	1	8.814
9	10.722	1	12.454	0	0
10	11.128	0	0	0	0
11	11.128	0	0	0	0
12	11.128	0	0	1	9.516
13	10.860	0	0	1	9.280
14	10.609	0	0	0	0
15	10.609	1	12.133	0	0
16	10.914	0	0	1	9.414
17	10.685	0	0	0	0
18	10.685	0	0	1	9.228
19	10.474	0	0	0	0
20	10.474	1	11.892	1	9.055
21	10.535	0	0	0	0
22	10.535	1	11.920	1	9.150
23	10.592	0	0	1	9.223
24	10.406	0	0	1	9.051
25	10.227	0	0	1	8.885
26	10.055				

TABLE 3. *Simulation Run 3*

n	X_n	U_n^+	V_n^+	U_n^-	V_n^-
1	9.000	0	0	1	6.000
2	8.400	0	0	0	0
3	8.400	1	10.680	1	6.120
4	8.660	1	10.782	1	6.539
5	8.870	1	10.876	1	6.863
6	9.046	0	0	0	0
7	9.046	1	10.890	1	7.202
8	9.183	0	0	1	7.399
9	8.927	0	0	0	0
10	8.927	0	0	0	0
11	8.927	0	0	1	7.279
12	8.712	0	0	1	7.101
13	8.512	1	10.092	0	0
14	8.785	0	0	0	0
15	8.785	1	10.310	1	7.261
16	8.862	0	0	0	0
17	8.862	0	0	0	0
18	8.862	1	10.318	1	7.405
19	8.929	1	10.365	1	7.492
20	8.993	0	0	0	0
21	8.993	1	10.394	0	0
22	9.213	0	0	0	0
23	9.213	0	0	0	0
24	9.213	1	10.568	1	7.857
25	9.266	0	0	0	0
26	9.266				

BIBLIOGRAPHY

[1] Dean, B. V., "Contract Award and Bidding Strategies," IEEE Transactions on Engineering Management *EM-12*, 53–59 (1965).

[2] Dupač, V., "On the Kiefer-Wolfowitz Approximation Method," Casopis Pest. Mat. *82*, 47–75 (1957). English translation in *Selected Translations in Mathematical Statistics and Probability*, Am. Math. Soc., Vol. IV (1963).

[3] Dvoretsky, A., "On Stochastic Approximation," Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability *1*, 39–55 (1956).

[4] Edelman, F., "Art and Science of Competitive Bidding," Harvard Business Review *43*. 53–66 (1965).

[5] Fabian, V., "Stochastic Approximation of Minima with Improved Asymptotic Speed," Ann. Math. Stat. *38*, 191–200 (1967).

[6] Fabian, V., "On Asymptotic Normality in Stochastic Approximation," Ann. Math. Stat. *39*, 1327–1332 (1968).

[7] Feller, W., *An Introduction to Probability Theory and its Applications* (John Wiley and Sons, Inc., N.Y.), Vol. I, 3rd Ed., 1968; Vol. II, 2nd Ed., 1971.

[8] Friedman, L., "A Competitive Bidding Strategy," Operations Research *4*, 104–112 (1956).

[9] Griesmer, J. H. and M. Shubik, "Toward a Study of Bidding Processes: Some Constant-Sum Games," Nav. Res. Log. Quart. *10*, 11–21 (1963).

[10] Griesmer, J. H. and M. Shubik, "Toward a Study of Bidding Processes, Part II: Games with Capacity Limitations," Nav. Res. Log. Quart. *10*, 151–173 (1963).

[11] Griesmer, J. H. and M. Shubik, "Toward a Study of Bidding Processes, Part III: Some Special Models," Nav. Res. Log. Quart. *10*, 199–217 (1963).

[12] Griesmer, J. H., R. E. Levitan, and M. Shubik, "Toward a Study of Bidding Processes, Part IV: Games with Unknown Costs," Nav. Res. Log. Quart. *14*, 415–433 (1967).

[13] Hanssmann, F. and B. H. P. Rivett, "Competitive Bidding," Oper. Res. Quart. *10*, 49–55 (1959).

[14] LaValle, I. H., "A Bayesian Approach to an Individual Player's Choice of Bid in Competitive Sealed Auctions," Management Science *13*, 584–597 (1967).

[15] Mercer, A. and J. I. T. Russell, "Recurrent Competitive Bidding," Oper. Res. Quart. *20*, 209–221 (1969).

[16] Reichert, A. O., "Models for Competitive Bidding under Uncertainty," Department of Operations Research, Stanford University, Technical Rept. No. 103 (1968). DDC No. AD663909.

[17] Schmetterer, L., "Stochastic Approximation," Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability *1*, 587–609 (1961).

[18] Sebly, S. M., Editor, ***CRC Standard Mathematical Tables***, The Chemical Rubber Co., Cleveland, (1969), 17th Ed.

[19] Stark, R. M. and R. H. Mayer, Jr., "Some Multi-Contract Decision-Theoretic Competitive Bidding Models," Operations Research *19*, 469–483 (1971).

[20] Stark, R. M., "Competitive Bidding: A Comprehensive Bibliography," Operations Research *19*, 484–490 (1971).

[21] Vickrey, W., "Counterspeculation, Auctions, and Competitive Sealed Tenders," Journal of Finance *16*, 8–37 (1961).

[22] Wasan, M. T., *Stochastic Approximation* (Cambridge University Press, 1969).

[23] Wilde, D. J., *Optimum Seeking Methods* (Prentice-Hall, Inc., Englewood Cliffs, N. J., 1964).

[24] Wilson, R. B., "Competitive Bidding with Asymmetric Information," Management Science *13*, 816–820 (1967).

STOCHASTIC DUELS INVOLVING RELIABILITY

David E. Thompson

Vector Research, Inc.
Ann Arbor, Michigan

ABSTRACT

The reliability of weapons in combat has been treated by Bhashyam in the context of a stochastic duel characterized by fixed ammunition supplies, negative exponentially distributed firing times and weapon lifetimes, and a fixed number of spare weapons for each duelist. The present paper takes a different approach by starting with the fundamental duel of Ancker and Williams, characterized by unlimited ammunition and by ordinary renewal firing times, and adding to it weapon lifetimes which can be functions of time or of round position in the firing sequence. Probabilities of winning and tieing are derived and it is shown that under certain conditions the weapon lifetimes are equivalent to random time and ammunition limits.

INTRODUCTION

The catastrophic failure of weapons during combat has been examined by Bhashyam [4] in the context of a one-versus-one stochastic duel characterized by the following restrictions:

(i) fixed ammunition supplies

(ii) negative exponentially distributed firing times

(iii) Erlang weapon lifetimes; or, equivalently, negative exponential lifetimes, and a fixed supply of spare weapons.

The author's work [5] differs from Bhashyam's by dropping the ammunition limitation in favor of including more general firing and failure time distributions. This paper starts with the fundamental duel of Ancker and Williams [3], characterized by unlimited ammunition and ordinary renewal firing times, and adds the feature of terminating each contestant's firing process after a random time interval or the expenditure of a random number of rounds.

While Bhashyam treated failed weapons as "sitting ducks," certain to be destroyed unless the opponent failed or ran out of ammunition, the present paper allows the instantaneous withdrawal of the failed weapon in addition to treating Bhashyam's case. It will be shown that the combination of time-dependent reliability with withdrawal has already appeared in the literature as a duel with a random time limit [2], and that the combination of round-dependent reliability with no withdrawal has appeared in the literature in the guise of a duel with random ammunition supplies [1].

TIME-DEPENDENT FAILURES

Given two opponents, A and B, the fundamental duel is defined by the following:

$f_A(t) =$ probability density of the time between A's rounds

$f_B(t) =$ probability density of the time between B's rounds

$p_A =$ single-shot kill probability for A versus B

$p_B =$ single-shot kill probability for B versus A.

This duel may be considered as a marksmanship contest in which the first contestant to kill a passive target wins. For weapons, which are completely reliable, we define the random variables:

T_A = time for A to kill his target, and

T_B = time for B to kill his target.

We now consider firepower lifetimes, which are functions of the time duration of the duel, and define

L_A = lifetime of A's weapon,

L_B = lifetime of B's weapon,

$r_A(t)$ = probability density of L_A,

$r_B(t)$ = probability density of L_B,

$R_A(t)$ = complementary distribution function of L_A, and

$R_B(t)$ = complementary distribution function of L_B.

When a failed weapon is allowed to withdraw instantaneously, the duel ends in a draw, if no kill occurs first. This duel possesses a time limit with distribution function $1-R_A(t)R_B(t)$. The win and tie probabilities for this random time-limited duel were derived by Ancker [2].

Assume now that the failed weapon remains in the duel until destroyed or the opponent fails. For completely reliable weapons the probability densities of T_A and T_B are, respectively,

(1) $$h_A(t)=\sum_{n=1}^{\infty} p_A q_A^{n-1} f_A^{*n}(t)$$

and

(2) $$h_B(t)=\sum_{n=1}^{\infty} p_B q_B^{n-1} f_B^{*n}(t),$$

where the asterisk denotes the n-fold convolution of a function with itself, $q_A=1-p_A$, and $q_B=1-p_B$. The complementary distribution function of T_B is denoted

(3) $$G_B(t)=\int_t^{\infty} h_B(x)\,dx,$$

and the complementary distribution function of T_A is similarly denoted $G_A(t)$. The probability A wins is given by

(4) $$P(A)=\int_0^{\infty} h_A(t)R_A(t)\left[1-\int_0^t h_B(x)R_B(x)\,dx\right]dt,$$

where the quantity in brackets is the probability B's target is alive at time "t." The tie probability is

(5) $$P(AB)=\int_0^{\infty} r_A(t)G_A(t)\,dt \times \int_0^{\infty} r_B(t)G_B(t)\,dt.$$

If the firing times are exponential and the failure times are Erlang-distributed, we obtain Bhashyam's duel for the case of an unlimited ammunition supply.

ROUND-DEPENDENT FAILURES

We now consider reliability as a function of weapon usage, as measured by the number of rounds fired. We define

$$\alpha_j = Pr[A \text{ fails on round } j+1]$$

$$\beta_k = Pr[B \text{ fails on round } k+1]\,,$$

where

$$\sum_{j=0}^{\infty} \alpha_j = \sum_{k=0}^{\infty} \beta_k = 1.$$

When a failed weapon is forced to remain in the duel, the situation is identical to that in which the two contestants start with random supplies of "j" and "k" rounds having distributions $\{\alpha_j\}$ and $\{\beta_k\}$. This duel with random ammunition supplies has already been treated by Ancker [1].

Ancker also treated a random ammunition-limited duel in which a participant withdraws immediately after firing his last allotted round. If one interprets a supply of "k" rounds as a weapon failure on round "$k+1$," the withdrawal cannot occur until one round later than in Ancker's ammunition-limited duel, on round "$k+1$" rather than round "k." The assumption is that in the reliability-limited duel a combatant does not know anything is wrong with his weapon until it fails to fire. For this duel the probability A kills his target in the interval $(t, t+dt)$ is $h_A(t)dt$, where

$$h_A(t) = \sum_{n=1}^{\infty} p_A q_A^{n-1} f_A^{*n}(t)\left(\sum_{j=n}^{\infty} \alpha_j\right). \tag{6}$$

At time "t" the probability B has neither killed his target nor withdrawn is

$$G_B(t) = \int_t^{\infty} f_B(x)dx + \sum_{n=1}^{\infty}\left(\sum_{k=n}^{\infty} \beta_k\right) q_B^n \int_0^t f_B^{*n}(x)\left[\int_{t-x}^{\infty} f_B(u)du\right]dx. \tag{7}$$

A similar definition holds for $G_A(t)$. The probability A wins is given by (8).

$$P(A) = \int_0^{\infty} h_A(t) G_B(t)dt. \tag{8}$$

The tie probability is given by (9).

$$P(AB) = Pr[L_A < L_B,\, L_A < T_A,\, L_A < T_B] + Pr[L_B < L_A,\, L_B < T_A,\, L_B < T_B].$$

$$P(AB) = \int_0^{\infty} G_B(t) w_A(t)dt + \int_0^{\infty} G_A(t) w_B(t)dt. \tag{9}$$

The quantities $w_A(t)dt$ and $w_B(t)dt$ are the probabilities of A and B, respectively, failing in the interval $(t, t+dt)$ without having killed their targets yet. The functions $w_A(t)$ and $w_B(t)$ are displayed in (10) and (11).

$$w_A(t) = \sum_{j=0}^{\infty} \alpha_j q_A^{j+1} f^{*\,j+1}(t). \tag{10}$$

$$w_B(t) = \sum_{k=0}^{\infty} \beta_k q_B^{k+1} f_B^{*k+1}(t). \tag{11}$$

ACKNOWLEDGMENTS

This paper is based upon the author's thesis [5] for his master's degree from the Department of Industrial Engineering at the University of Michigan; the writing of the thesis was supported, in part, by the Office of Naval Research and WSAD. Special thanks are extended to Professor Seth Bonder for his help and guidance.

REFERENCES

[1] Ancker, C. J., Jr., "Stochastic Duels with Limited Ammunition Supply," Operations Research *12*, 38–50 (1964).

[2] Ancker, C. J., Jr., "Stochastic Duels of Limited Time-Duration," J. Canadian Operational Research Society *4*, 69–81 (1966).

[3] Ancker, C. J., Jr., and T. Williams, "Stochastic Duels," Operations Research *11*, 803–817 (1963).

[4] Bhashyam, N., "Stochastic Duels with Nonrepairable Weapons," Nav. Res. Log. Quart. *17*, 121–129 (1970).

[5] Thompson, D. E., "Mobility and Reliability in the Theory of Stochastic Duels," unpublished master's thesis, Dept. of Industrial Engineering, The University of Michigan, Ann Arbor, Michigan (Aug. 1968).

THE LAW OF AVERAGES AS A COMPUTING TOOL

L. E. N. Delbrouck

Bell-Northern Research
Ottawa, Ontario

ABSTRACT

This paper is concerned with the description of a computational technique designed to formulate and then calculate delay probabilities as long run averages of series of independent statistical trials. Specific examples illustrating the effectiveness of the method are also given.

INTRODUCTION

A hybrid compute-simulate method especially designed for the evaluations of delay distributions in the M/G/1 queue is described in this paper. The method is based on the old idea of approximating volumes by generating random numbers and taking a certain well-defined course of action depending on whether or not those numbers fall within a specified region. More precisely, let X be a random variable defined over a sample space S and such that for every Borel set B on the real line, the event $\{X\epsilon B\}$ is assigned the probability $Q(B)$. A probability measure P is defined on S, and for $P-$ almost all $s\epsilon S$, the conditional probability of the event $\{X\epsilon B\}$ given s is assumed to exist, and it is designated $Q(B/s)$. By definition

$$Q(B) = \int Q(B/s)\, P(ds)$$

$$= E[Q(B|s)],$$

where E is the expected value operator.

Let us now draw at random and independently, according to P, N sample points, $s_1, s_2, \ldots, s_N$ from S and then calculate the probabilities $Q(B/s_1)$, $Q(B/s_2)$, $\ldots$ $Q(B/s_N)$. By the weak law of large numbers, we know that for any positive ϵ no matter how small:

$$\lim_{N\to\infty} \text{Prob}\left(\left|Q(B) - N^{-1}\sum_{i=1}^{N} Q(B/s_i)\right| < \epsilon\right) = 1.$$

In other words, for large values of N, we are practically certain that the arithmetic mean,

$$N^{-1}\sum_{i=1}^{N} Q(B/s_i),$$

is very close to $Q(B)$.

By means of a simple example, we will exhibit the mechanism of a practical method of computation

based on these notions. Thereafter, we will consider the more difficult problem of evaluating cumulative delay distributions when the order of service is last-come first-served (LIFO) or selection at random (SIRO). Subsidiarily, we will also consider the busy period distribution.

The research leading up to this article was primarily motivated by the need for a workable method to approximate delay—SIRO distributions, for single server systems with Poisson arrivals and discrete service time distributions. Typical of such systems is the facility which handles several classes of Poisson distributed requests, with demands for service in each class being allotted a constant service time which, however, varies from class to class. The transforms of such delay distributions are quite complex (see for example [4] and [8]) and do not appear to yield readily to numerical inversion. Indeed Burke [2] who considered the case of a single class of requests did not use transforms, and Riordan [6] who treated the case of the exponential distribution of service time used the special properties of systems of difference-differential equations with tridiagonal coefficient matrices. By contrast the approach herein discussed recommends itself for several reasons. It circumvents, on the one hand, the serious difficulties inherent to a strictly analytical treatment, and on the other it avoids the nuisance of correlated outputs, onset of stationary conditions and constant updating of the system that besets an all-out simulation effort.

1. A SIMPLE EXAMPLE

Consider the M/M/1 queue for which the mean inter-arrival time is λ^{-1} and the mean service time 1. It is assumed that $\lambda < 1$, that the queue is in the steady-state, and that the queue discipline is first-come-first-served (FIFO). A new arrival finds $n(\geqslant 0)$ calls in the system with probability $\Pi_n = (1-\lambda)\lambda^n$ and its actual delay (time of waiting in line) is then the sum of n exponentially distributed values with mean 1. If this sum is at least t, then obviously the delay cannot be less than t; otherwise it must be less than t. In other words, if for $t(\geqslant 0)$ held fixed, we let $P(t)$ denote the steady-state probability that the actual delay of an arbitrarily chosen call is not less than t, and if by $P(t/n, \sigma_n)$ we designate the conditional version of this probability given that the call encounters n other calls upon arrival and that the sum of n service times is σ_n; then

$$(1.1) \qquad P(t/n, \sigma_n) = 1, \quad \text{if } \sigma_n \geqslant t, \qquad = 0, \quad \text{otherwise.}$$

On the basis of (1.1), we will organize a game for which the long run average is $P(t)$. First draw an integer, say k, according to the geometric distribution $\{\Pi_n : n \geqslant 0\}$. Next, draw k samples $t_1, t_2, \ldots, t_k$ from an exponential population with mean 1. If $t_1 + t_2 + \ldots + t_k \geqslant t$, we score 1; otherwise the score is 0. We repeat this game N times, sum the score, and divide the total through N thereby obtaining the average score per game. This average is an unbiased estimate of $P(t)$, and for N large, we expect it to be a good approximation to the probability in question.

2. BASIC QUEUEING QUANTITIES ON THE M/G/1 QUEUE

We summarize certain well-known facts regarding the queue herein considered (see for example [7]). The inter-arrival times are again negative exponential with mean λ^{-1}. The service time distribution is denoted H, and

$$\alpha = \int_0^\infty \{1 - H(z)\}\, dz < \infty.$$

Equilibrium conditions are understood to hold; i.e., $\lambda\alpha < 1$. For $n = 0, 1, 2 \ldots$, the steady-state probability that there are n calls in the system is denoted Π_n. The sequence $\{\Pi_n : n \geqslant 0\}$ satisfies the recursive relations

$$\Pi_0 = 1 - \lambda\alpha,$$

$$\Pi_1 = \Pi_0 a_0^{-1}(1 - a_0),$$

$$\Pi_{n+1} = a_0^{-1}\{(1-a_1)\Pi_n - a_2\Pi_{n-1} \ldots (\Pi_0 + \Pi_1)a_n\},\ n \geqslant 1, \tag{2.1}$$

where

$$a_x = \int \frac{(\lambda s)^x}{x!} e^{-\lambda s} dH(s),\ x = 0, 1, 2, \ldots. \tag{2.2}$$

For $t \geqslant 0$, otherwise fixed, let $P(t)$ denote the probability that in the steady-state the actual delay of any call exceeds t. By $W(t)$, we mean the probability of the same event for a delayed call. Hence

$$W(0) = 1$$

and

$$P(t) = \lambda\alpha W(t),\ t > 0. \tag{2.3}$$

We are interested in evaluating $W(t)$ for any $t > 0$, and hence we will confine our attention solely to calls that are delayed.

Consider such a call arriving at an instant 0. By $N(t)$, we mean the number of service periods beginning in the interval $(0,t)$. For $i = 1, 2, \ldots, N(t)$, we let τ_i, t_i, $x(t_i)$ denote, respectively, the initial moment of the ith such period, its duration, and the number of new arrivals therein. For the sake of consistency, we let t_0, $x(t_0)$ represent the length of and number of arrivals in the interval $(0,\tau_1)$.

For $i \geqslant 1$, the variables, t_i, are independent and distributed according to H; they are also independent of the variable t_0 which has the distribution H* specified by the relation.

$$\text{Prob}\ (t_0 \leqslant x) = H^*(x) = \alpha^{-1} \int_0^x \{1 - H(z)\}dz,\ x > 0. \tag{2.4}$$

To avoid notational complexities and cumbersome circumlocution, let us agree that the relation $x \sim X$ signifies that x is a particular value of a variable X. For the same reason, the symbol P_μ is used to characterize the Poisson arrival distribution with mean μ per unit time. In particular, the variable $x(t_i)$ has the distribution $P_{\lambda t_i}$. Occasionally, we will also abbreviate $x(t_i)$ to x_i.

Depending on the queue discipline, we shall be led to construct a probability space $S(t)$ for the sample realizations of the joint process regulating the flow of arrivals to and departures from the system throughout the interval $(0,t)$. We shall then proceed to "draw" at random, according to some suitably defined probability measure P, N independent samples, $s_1, s_2, \ldots, s_N$ from $S(t)$. For $i=1, 2, \ldots N$ the conditional probability $W(t/s_i)$, given the realization s_i, that the delayed call fails to obtain service in $(0,t_i)$ will then be computed, and the mean

$$\tilde{W}(t) = N^{-1} \sum_{i=1}^{N} W(t/s_i)$$

will be taken as an approximation to $W(t)$.

We need deal but briefly with the steady-state distribution of the delay-FIFO. Indeed in this case (see for example [8]), inversion of the Pollaczek-Khintchine formula shows that the delay may be viewed as a sum of independent variates, each with distribution H^*, and the number of terms in this sum has the geometric distribution with parameter $\lambda\alpha$. Hence the delay distributions may be modelled in a fashion entirely similar to that outlined in section 1. However, should moments of this distribution be easily accessible, a straightforward Hermite-Chebyshev approximation or similar expansion would probably be more satisfactory.

In the next section we shall consider the case of the delay-LIFO and the busy period. The delay SIRO will be examined in section 4.

3. THE DELAY DISTRIBUTION WHEN THE DISCIPLINE IS LIFO

The space $S(t)$ consists of all the possible realizations of the processes $\{t_0 : t_0 > t\}$ and $\{(t_i, x(t_i)) : i=0, 1, 2 \ldots, N(t)-1\}$. Typically, any element s in $S(t)$ is of the form $\{\theta_0\}$ and it satisfies

$$\theta_0 \sim t_0,$$

(3.1) $$\theta_0 > t;$$

or it is of the form $\{\theta_0, x_0^*), (\theta_1, x_1^*), \ldots, (\theta_{m-1}, x_{m-1}^*)\}$, where

$$\theta_i \sim t_i,$$

(3.2) $$x_i^* \sim x(\theta_i), \; i=0, 1, 2, \ldots, m-1,$$

and the integer m satisfies the condition

(3.3) $$\tau_m \leq t < \tau_{m+1}.$$

Let us now define $C(t)$ to be the subset of $S(t)$ consisting of all elements of form (3.1), together with those elements of form (3.2) satisfying the conditions

$$x_0^* \geq 1,$$

$$x_0^* + x_1^* \geqslant 2,$$

and

(3.4) $$\chi_0^* + x_1^* \ . \ . \ . + x_{N(t)-1} \geqslant N(t).$$

It should be clear that the aggregate $C(t)$ represents the event "throughout the interval $(0, t)$, the service channel is continuously busy, or when discharging a call, it is immediately seized by a call that arrived after time 0." Therefore, given any element $s \in S(t)$, according to whether s does or does not belong to $C(t)$, the call arriving at 0 does or does not fail to obtain service in $(0, t)$, i.e.,

$$W(t/s) = 1, \qquad \text{if } s \in C(t)$$

$$= 0, \qquad \text{otherwise.}$$

To evaluate $W(t)$, we may therefore design a somewhat more elaborate version of the binomial game briefly outlined in the first section of this paper. A game is carried out in the following manner. First draw $\theta_0 \sim t_o$ and if $\theta_0 > t$, score 0 and repeat the game. In general, a game will consist of generating pairs (θ_0, x_0^*), (θ_1, x_1^*), . . ., such that $\theta_i \sim t_i$, $x_1^* \sim x(t_i)$, until the sum $\theta_0 + \theta_1 + \theta_2$ . . . finally overshoots t, for which we score 1, or until the chain of conditions

$$\{x_0^* \geqslant 1\} \cap \{x_0^* + x_1^* \geqslant 2\} \cap \{x_0^* + x_1^* + x_2^* \geqslant 3\} \ . \ . \ .$$

is finally broken, in which case, we score 0. The total score divided by the number of games is then our approximation for $W(t)$.

For automatic computation (see section 5), it will be simpler to generate inter-arrival times ω_0, ω_1, ω_2, . . ., independently and award a score of 1 provided the chain of conditions $\theta_0 > \omega_0$, $\theta_0 + \theta_1 > \omega_0 + \omega_1$, $\theta_0 + \theta_1 + \theta_2 > \omega_0 + \omega_1 + \omega_2$, . . . remains unbroken, or until the sum $\theta_0 + \theta_1 + \theta_2 +$ . . . finally exceeds t.

In closing this section, let it be noted that if we specify the distributions of t_i to be H rather than H^* then the procedure discussed here yields an approximation to the probability that the server's busy period exceeds t. It is to be emphasized incidentally that insofar as delays are concerned the memoryless property of the Poisson input enables us to ignore the joint distribution of the state of the system and remaining service time at points of arrival. This may also be valid for the queue with arbitrary renewal input and exponential service time, a case which was not considered in this study. But it would no longer hold true for the general queue GI/G/1. Since, however, the busy period begins with the incipience of a service time following an idle period, estimation of busy period distributions through the approach herein considered is valid for the general queue as well.

4. THE DELAY WHEN THE DISCIPLINE IS SIRO

When the random service discipline prevails, a somewhat more elaborate simulation scheme is required. To begin with, the sample functions $W(t/s)$ are no longer of the zero-one variety. Moreover one must determine the queue size at $\tau_1 - 0$; and this will depend on the residual service time, t_0. To obviate this difficulty we let ν denote the length of the service time during which the new caller arrives,

so that (see Reference [4])

$$P(\nu \leqslant x) = F(x) = \frac{\int_0^x z\,dH(z)}{\int_0^\infty z\,dH(z)}. \tag{4.1}$$

Furthermore, we set

$$a(x, z) = \frac{(\lambda z)^x}{x!} e^{-\lambda z}; \qquad z \geqslant 0,\ x = 0, 1, 2, \ldots . \tag{4.2}$$

Conditioned on ν, the residual service time t_0 is now designated $t_0(\nu)$ and it is uniformly distributed on the interval $(0, \nu)$. Again, the queue size at $\tau_1 - 0$ is also dependent on ν, and accordingly, it is denoted $n_1(\nu)$. Its conditional distribution given ν satisfies (see Reference [2])

$$P(n_1(\nu) = n) = (\Pi_0 + \Pi_1)\, a(n-1, \nu) + \Pi_2\, a(n-2, \nu) + \ldots + \Pi_n\, a(0, \nu). \tag{4.3}$$

The sample space $S(t)$ will now consist of the elements $\{\theta_0\}$ such that $\theta_0 \sim t_0(\nu)$, and of the elements $\{(\theta_0, n^*), (\theta_1, x_1^*), \ldots, (\theta_{N(t)-1}, x^*_{N(t)-1})\}$, such that for $i = 1, 2, \ldots, N(t)-1$, the pairs (θ_i, x_i^*) are obtained as in section 3. To generate random pairs (θ_0, n^*), one may proceed as follows:

STEP 1: Generate a sample $\nu^* \sim \nu$,

STEP 2: Generate a uniform random number R, and calculate $\theta_0 = R \times \nu^*$,

STEP 3: Draw an integer i according to the distribution $\{\Pi_n : n \geqslant 0\}$,

STEP 4: Draw an integer j according to the Poisson distribution $\{a(n, \nu^*) : n \geqslant 0\}$,

STEP 5: Calculate $n^* = i + j + 1$.

Now for a sample value $s = \theta_0 > t$, we have

$$W(t/s) = 1. \tag{4.4}$$

In this case, only Steps 1 and 2 are required. If $s = \{(\theta_0, n^*), (\theta_1, x_i^*), \ldots, (\theta_{N(t)-1}, x^*_{N(t)-1})\}$, then we set:

$$W(t/s) = \frac{n^* - 1}{n^*} \prod_{i=1}^{N(t)-1} \frac{\left(n^* + \sum_{j=1}^{i} x_j^* - (i+1)\right)}{\left(n^* + \sum_{j=1}^{i} x_j^* - i\right)}. \tag{4.5}$$

Again, the "game" continues until the sum $\theta_0 + \theta_1 + \theta_2, \ldots$ overshoots t, and then the score is computed according to (4.4) or (4.5), or until the cumulative product $\left\{\frac{n^*-1}{n^*}\right\}\left\{\frac{n^* + x_1^* - 2}{n^* + x_1^* - 1}\right\}\left\{\frac{n^* + x_1^* + x_2^* - 3}{n^* + x_1^* + x_2^* - 2}\right\}$ $\ldots$ becomes 0. In this case the score is zero.

5. TEST RESULTS

The validity of the approach described here was tested by means of two Fortran programs called XPLCFS and XPROS. The first one provides for the approximation of the delay probability $W(t)$ in

the queue with constant service time of unit length and queue discipline LIFO. For the second program, the discipline is SIRO, the service time negative exponential with mean 1. Both programs use a pseudo-uniform random number generating subroutine. The exponential inter-arrival or service time is obtained by generating a random number and applying the logarithmic transformation suitably scaled. The geometric queue size is simulated by generating a random number applying the logarithmic transformation and then discarding the fractional part. The results for the delay SIRO are compared with Riordan's [6] and Kingman's [4] approximations. For the delay-LIFO the values are contrasted to those obtained through the formula [8]:

$$W(x) = \frac{1 - \left\{(1-\lambda x) + \lambda \sum_{j=1}^{\infty} e^{-\lambda x} \frac{(\lambda x)^{j-1}}{j-1!} (\min \{x, j\})\right\}}{\lambda \alpha}.$$

The number of replicates N, as well as the computer time, including compilation on the CDC 3300 computer are also quoted for several values of the probabilities.

For completeness, we also quote two results obtained by means of a short program XPFIFO prepared to test the procedure outlined in Section 1. In this case the exact value of the probability $P(t)$ is $\lambda e^{-t(1-\lambda)}$.

The results for these preliminary tests are collected in Tables A–1, A–2, and A–3. On the basis of these results, several programs were written to calculate queueing quantities for the Poisson queue

TABLE A–1

λ	t	$P(t)$			N	Computing Time
		Kingman	Riordan	XPROS		
0.9	4	0.498	0.500	0.455	5,000	58″
0.9	10	0.280	0.280	0.266	10,000	3′ 7″
0.9	20	0.140	0.130	0.134	3,000	1′ 56″
0.9	50	0.033	0.028	0.034	1,500	2′ 22″
0.7	5		0.129	0.125	10,000	1′ 18″
0.7	10		0.0466	0.0455	10,000	1′ 49″
0.45	5		0.0377	0.0202	10,000	23″
0.45	10		0.00668	0.00274	10,000	50″

TABLE A–2

λ	t	$W(t)$		N	Computer Time
		Formula	XPLCFS		
0.95	10	0.104	0.106	10,000	50″
0.95	20	0.0673	0.0674	5,000	40″
0.9	10	0.086	0.0847	10,000	51″
0.6	5	0.0447	0.0348	5,000	20″
0.5	5	0.0248	0.0189	10,000	

TABLE A–3

λ	t	$P(t)$		N	Computer Time
		Formula	XPFIFO		
0.9	8	0.3944	0.3937	10,000	
0.9	10	0.3312	0.3265	10,000	60″

with discrete service time distribution [3]. The largest such program called STATROS is designed to evaluate recursively the steady-state probabilities $\{\Pi_n : n \geqslant 0\}$ up to a maximum M either specified beforehand or such that the quantity $1-\sum_{j\leq M} \Pi_j$ is less than a small preassigned value; the program uses these probabilities to generate queue sizes in the estimation of the probability $W(t)$. The case of the one step (constant) service time was used for testing purposes. Here the value $\lambda=0.8$ was kept fixed, while t-values ranged between 3 and 50; 5,000 replications were made in each case. The probabilities $W(t)$ obtained thereby were then used to calculate $P(t)=\lambda\alpha W(t)$. The latter values were then compared with the same probabilities read-off beforehand from the corresponding curves in Burke's paper ([2], Figures 2 and 3, pp. 1026, 1028). The results of this test are given in Table B–1.

TABLE B–1

Time $=t$	Value of $P(t)$ (STATROS)	Estimated theoretical value of $P(t)$	Computer time
3.00	0.196	0.195	
3.25	0.184	0.182	
3.50	0.171	0.169	
3.75	0.157	0.156	4′ 34″
4.00	0.151	0.142	
4.25	0.137	0.134	
4.50	0.126	0.127	
4.75	0.118	0.119	
5.00	0.1065	0.112	
5.25	0.1025	0.105	
5.50	0.0975	0.098	
5.75	0.0925	0.092	5′ 55″
6.00	0.085	0.085	
6.30	0.076	0.078	
6.70	0.0705	0.069	
6.95	0.0665	0.064	
7.00	0.064	0.063	
8.00	0.052	0.050	
9.00	0.0425	0.042	
10.00	0.033	0.033	8′ 07″
11.00	0.027	0.028	
12.00	0.021	0.022	
13.00	0.019	0.017	
14.00	0.0153	0.016	
20.00	0.0062	0.0055	
30.00	0.0015	0.0014	8′ 01″
40.00	0.00063	0.00045	
50.00	0.000146	0.00013	

It will be noted that on the whole the compute-simulate approach yields valid results, there being sometimes a tendency, however, for the small probabilities of rare events to be underestimated. This is to be expected of any method based on simulation. More difficult to explain, however, is the fact that this tendency should be quite noticeable in Table A–1, less pronounced in Table A–2, and marginal only in Table B–1. The writer did not have the opportunity to investigate this peculiarity and for the time being can only ascribe it to a combination of factors, such as differences in queueing disciplines and variabilities of service times. In the first instance, however, it has been pointed out [4] that Riordan's approximation may not be uniformly accurate although ironically perhaps the region where it should break down—large delays or heavy traffic—is that in which agreement with our results is good. In any case, it is quite possible that this approximation is not altogether reliable for small probability values, and the discrepancy noted at the bottom of Table A–1 may not be quite so large in actual fact.

6. ON STATISTICAL INFERENCE AND THE CONSTRUCTION OF EMPIRICAL DISTRIBUTIONS

The method described in the foregoing sections yields estimates the variability of which may be assessed by means of standard techniques. As an example, in the case of the delay-LIFO the N replications constitute a sequence of N Bernoullian trials each of which ends in success with probability $W(t)$. For large N the Student-t or normal distributions may be used to draw the proper conclusion regarding the true value $W(t)$. Even though the procedure is not binomial when the discipline is SIRO, the central limit theory still applies.

This method may also be used to construct empirical distributions. For example, referring to the delay-LIFO, suppose that we set $t=\infty$, and instead of scoring we record the length d elapsing between time 0 and the instant when condition (3.4) is finally broken. Letting d_i denote that length observed in the ith replication of the experiment, we note that the sequence $d_1, d_2, \ldots d_N$ is a sample of N independent delay values each of which is drawn according to the distribution $1-W(\quad)$, i.e.,

$$P(d_i \leq x)=1-W(x);\ i=1, 2, \ldots, N.$$

This sample satisfies the premises of the Glivenko-Cantelli Lemma (see [5] p. 20), and therefore, it may be used to construct an empirical distribution. Since the distribution $1-W(\quad)$ is clearly continuous, the Smirnov-Kolmogorov statistic is applicable.

At the expense of minor complications, not touched upon here, the same purpose may be achieved in the case when the discipline is SIRO. For the busy period, however, consideration should be given to the possibility that the distribution may not be continuous; the χ^2 statistic with $N-1$ degrees of freedom would be more appropriate.

7. CLOSING REMARKS

In the applications of queueing theory, as indeed in several other contexts, simulation is often used to approximate quantities that are all but inaccessible through brute force computation. Numerical inversion may provide an attractive alternative [1], but it cannot be wholeheartedly recommended unless the transform itself can be evaluated to a very high degree of accuracy. Comparatively, the method described here is numerically accessible, and it is also stable. In contrast to the more common system simulation approach whereby delays are obtained by clocking arrival and departure times, this method is not plagued by the vexing problem of correlated delays and onset of stationary conditions.

By treating $W(t)$ as the long range average of independent replicates of a statistical experiment, we may avail ourselves of the standard techniques to assess the goodness of the approximation.

The limitations on the method should, of course, be realized. To the extent that large quantities of uncorrelated and uniformly distributed random numbers may be generated, and accurately transformed as required, the method is in fact foolproof. Unavoidable departures from these ideal conditions do have an effect on the estimates obtained. For approximating small values of $W(t)$, the method may become erratic. This would be manifest from the "wiggly" tail of the delay curve being estimated and this portion of the curve would require retrials with larger samples of random numbers.

REFERENCES

[1] R. E. Bellman, et al., *Numerical Inversion of the Laplace Transform* (Elsevier, New York, 1966).

[2] P. J. Burke, "Equilibrium Delay Distributing for One Channel With Constant Holding Time, Poisson Input and Random Service," Bell System Technical Journal *38*, 1021–1031 (1959).

[3] L. E. N. Delbrouck, "The Poisson Queue With Discrete Constant Holding Time," Northern Electric Co., Ltd., R & D, Ottawa, Internal Report TM 8325–6–68 (1969).

[4] J. F. C. Kingman, "On Queues in Which Customers are Served in Random Order," Proc. Cambridge Phil. Soc. *58*, 79–91 (1962).

[5] M. Loève, *Probability Theory* (Van Nostrand, New York, 1960).

[6] J. Riordan, "Delay Curves for Calls Served at Random," Bell System Technical Journal *32*, 100–119 (1953).

[7] T. Saaty, Elements of Queuing Theory (McGraw-Hill, New York, 1961).

[8] L. Takács, "Delay Distributions for One Line With Poisson Input, General Holding Times, and Various Orders of Service," Bell System Technical Journal *42*, 487–503 (1963).

PREDICTION WITH ZERO-ONE LOSS STRUCTURE

Paul D. Berger

Boston University

This paper poses a prediction problem in which a linear model is assumed. With a "zero-one" loss structure as the loss from incorrect prediction, it is suggested that least squares may not be appropriate for estimating the parameters of the model. An alternate criterion is proposed and integer programming is used in order to find the estimates, given the proposed criterion.

1. INTRODUCTION

Let us suppose

$$\tilde{y}_j = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ij} + \tilde{\epsilon}_j, \tag{1}$$

and the β_i, $i=0, \ldots, k$, are known. If the density of $\tilde{\epsilon}_j/\underline{x}_j$ is known ($\underline{x}_j = x_{1j}, \ldots, x_{kj}$), a point estimate of $\tilde{y}_j/\underline{x}_j$ is found by minimizing the expected loss suffered from prediction. That is, if $\hat{y}_j$ is the prediction and the loss suffered is $l(\tilde{y}_j, \hat{y}_j)$, $\hat{y}_j$ is chosen to minimize

$$\int_{-\infty}^{\infty} l(y_j, \hat{y}_j)\, dF(y_j/\underline{x}_j), \tag{2}$$

where $F(y_j/\underline{x}_j)$ is the distribution function of $\tilde{y}_j$ given $\underline{x}_j$. (Since $F(\epsilon_j/\underline{x}_j)$ is known, $F(y_j/\underline{x}_j)$ is essentially known.)

In general, however, $\underline{\beta} = (\beta_0, \ldots, \beta_k)$ is not known and frequently $F(\epsilon_j/\underline{x}_j)$ is also unknown. In this situation an optimal prediction of $\tilde{y}_j$ is not so straight forward.

2. THE PREDICTION PROBLEM

Suppose we have the model in (1) with $\underline{\beta}$ unknown and $F(\epsilon_j/\underline{x}_j)$ also unknown. We have sample data consisting of n points $(y, \underline{x})$ and it is required to predict y_j for a given $\underline{x}_j$ (generally not part of our sample data) by estimating $\underline{\beta}$ by $\underline{b}$ and choosing as our prediction $\hat{y}_j$,

$$\hat{y}_j = b_0 + \sum_{i=1}^{k} b_i x_{ij}. \tag{3}$$

The first question posed is how to estimate $\underline{b}$. Least Squares is universally used for this purpose (although occasionally linear or perpendicular distances are minimized). If it is desired to minimize expected squared prediction error, then least squares may be appropriate. However, consider a situation in which the loss of prediction is

$$l(\tilde{y}_j, \hat{y}_j) = \begin{cases} K \text{ if } |y_j - \hat{y}_j| > \delta \\ 0 \text{ if } |y_j - \hat{y}_j| \leq \delta \end{cases}$$

This is a "zero-one" loss structure in which you either "win" or "lose." Here, least squares may be totally inappropriate. What could be appropriate is to find $\underline{b}$ such that of the n sample points, as many as possible satisfy

$$|y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| \leq \delta. \tag{4}$$

In general, $\underline{b}$ is not unique and many different ways could be proposed for "breaking the tie," one being finding the $\underline{b}$ which yields the same number of data points satisfying (4) with as small a $\delta_0 (\delta_0 \leq \delta)$ as possible.

One circumstance in which this criterion may yield a very different estimate from least squares is one in which $\tilde{\epsilon}_j$ consists of a mixture of two random variables; first $\tilde{\epsilon}_{1j}$ with mean 0 and variance σ^2, perhaps normally distributed, and $\tilde{\epsilon}_{2j} = \tilde{\epsilon}_{1j} + \tilde{J}$, where $\tilde{J}$ is a random variable with mean $\lambda \gg \sigma$. (J may stand for "Jolt"). We have $\tilde{\epsilon}_j = \tilde{\epsilon}_{1j}$ with probability $p(\underline{x})$ (unknown) and $\tilde{\epsilon}_j = \tilde{\epsilon}_{2j}$ with probability $1 - p(\underline{x})$. This situation arises often when J is the result of a specific factor which influences y_j, but we have not been able to identify this factor. (If we could identify it, we would include it in our model, perhaps, if appropriate, as a dummy variable.)

Examples of this type problem structure arise in many different contexts. In general, with spatial coordinates, one might wish to predict the straight path that is within a given range of the object. In medicinal applications one must decide upon the amount of drug to administer; for instance the amount of immunosuppressive drug to administer after a transplant operation in a situation where too much drug could cause infection and probable death, while too little will lead to rejection of the new organ and probable death. The correct dosage is likely a function of weight, sex, stamina, age, etc., as well as other unknown factors.

3. A NUMERICAL EXAMPLE

Consider the following sample data with which to estimate $\underline{\beta}$ in $\tilde{y}_j = \beta_0 + \beta_1 x_{1j} + \tilde{\epsilon}_j$.

x	y
1	0.9
1	1.1
1	5.0
1	1.0
2	2.1
2	1.9
2	2.0
2	8.0
3	3.0
3	8.0
3	2.9
3	3.1

This data is pictured in Figure 1.

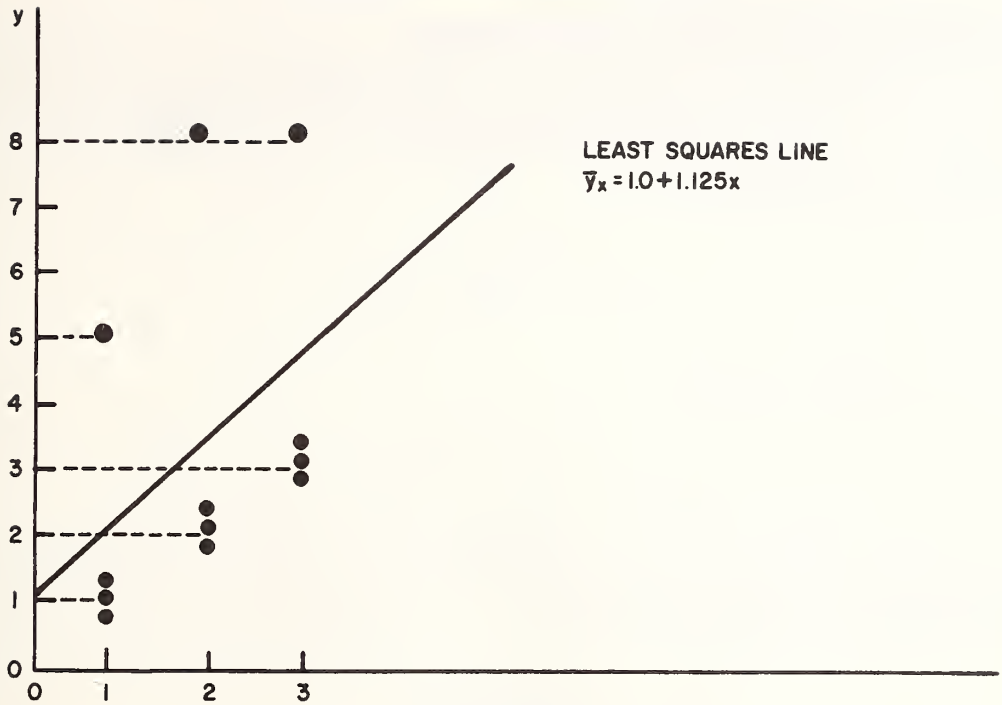


FIGURE 1.

The least squares line is calculated to be

$$\bar{y}_x = 1.0 + 1.125x,$$

and is imposed on Figure 1. If we assign δ to be, say, 0.5, let us compare $\bar{y}_x$ as a prediction against $\hat{y} = x$, which yields a prediction such that as many of the 12 points as possible (9 of them) satisfies $|\hat{y}_j - y_j| \leq \delta$. (Considering the secondary criterior of choosing among the nonunique $\underset{\sim}{b}$'s by successively reducing δ, keeping, in this case, nine points such that $|\hat{y}_j - y_j| \leq \delta_0$, we find at $\delta_0 = 0.1$, the above line is unique).

x	$\|\bar{y}_x - y\|$ with least squares	$\|\hat{y} - y\|$ with $\hat{y} = x$
1	1.225	0.1
1	1.025	0.1
1	2.875	4.0
1	1.125	0.0
2	1.150	0.1
2	1.350	0.1
2	1.250	0.0
2	4.750	6.0
3	1.375	0.0
3	3.625	5.0
3	1.475	0.1
3	1.275	0.1
Sums of *squared* deviations	$\|\bar{y}_x - y\|^2 = 58.41$	$\|\hat{y} - y\|^2 = 77.06$
Number of deviations within $\delta = 0.5$	0 of 12	9 of 12

It seems clear that $\underset{\sim}{b} = (0, 1)$ is better than $\underset{\sim}{b} = (1.0, 1.125)$ <u>here</u> (with a zero-one loss structure).

4. ALGORITHM FOR FINDING THE ESTIMATE

The object of this section is to formulate as a standard integer programming problem (by "standard," we refer to a standard linear programming formulation with the addition of specifications that certain variables be restricted to integer values) the finding of $\underline{b}$ such that as many points as possible out of n given $(y,\underline{x})$ points satisfy

$$|y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| \leqslant \delta.$$

Let Δ_j be defined for each j by

$$\Delta_j = \begin{cases} 1 & \text{if } |y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| \leqslant \delta \\ 0 & \text{otherwise.} \end{cases}$$

Then the problem may be stated to choose $\underline{b}$ to

(5) $$\text{maximize} \sum_{j=1}^{n} \Delta_j,$$

subject to, for each j,

(6a) $$\Delta_j \leqslant 1$$

(6b) $$\Delta_j \geqslant 0$$

(6c) $$\Delta_j \text{ is an integer}$$

(7) $$\Delta_j = 1 \quad \text{if and only if} \quad |y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| \leqslant \delta.$$

We now replace (7) by either

(8) $$|y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| \leqslant \delta,$$

or

(9) $$\Delta_j = 0.$$

It is important to note that (8) and (9) do capture the essence of (7), since if (8) holds, the *algorithm will choose* $\Delta_j = 1$, since (5) is maximizing an increasing function of Δ_j (i.e., the coefficient in the objective

function is positive).

We can in turn replace (8) and (9) by the set of constraints (10), (11), (12) as follows:

(10) $$|y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij}| - A_j M \leq \delta,$$

(11) $$\Delta_j + A_j M \leq M,$$

(12a) $$A_j \geq 0,$$

(12b) $$A_j \leq 1, \text{ and}$$

(12c) $$A_j \text{ is an integer.}$$

Where M is a "very large" number. Note that if $A_j = 0$, (10) is equivalent to (8) and (9) essentially disappears, while if $A_j = 1$, (11) is equivalent to (9) and (8) essentially disappears. This leaves constraints (6), (10), (11), and (12). We can remove the constraint involving absolute values by replacing (10) by the set

(13) $$y_j - b_0 - \sum_{i=1}^{k} b_i x_{ij} - A_j M \leq \delta$$

and

(14) $$b_0 + \sum_{i=1}^{k} b_i x_{ij} - y_j - A_j M \leq \delta.$$

The variables $b_0 \ldots, b_k$ are unrestricted in sign, but could in the typical way be transformed to positive variables by defining $b_i = b_{1i} - b_{2i}$ for each i and requiring b_{1i} and b_{2i} to be positive.

We thus have for each j constraints, (6), (11), (12), (13), (14) to satisfy in maximizing (5). Similarly, this may be viewed as constraints (11), (13), and (14) with Δ_j either 0 or 1, and A_j either 0 or 1. Then for n data points and k independent variables, we have $2n + (k+1)$ variables and $3n$ constraints in our integer programming problem with $2n$ variables restricted to 0 or 1.

5. DISCUSSION AND SUMMARY

We have discussed prediction problems when the loss structure is zero-one and concluded that may be appropriate to estimate the regression parameters not by the least squares criterion, but by a criterion of maximizing the number of data points within a preassigned distance (along the y axis) from the estimated equation. This procedure seems intuitively appealing and seems likely to be a good heuristic procedure for finding superior estimates when the loss of prediction is dependent only on whether we "hit" or "miss." The procedure seems appealing also from a Bayesian point of view in circumstances where the prior density is "gentle" and swamped by an unknown likelihood function. Work has been done to try in a reasonable way to estimate parameters by using procedures which are relatively insensitive to maverick data points; for instance, to use instead of the sample mean, the

arithmetic mean after throwing out the largest and smallest observations, or by computing the average of only the middle 50 percent of the data. If this type procedure is desired, it seems likely that this regression estimation criterion could be used heuristically to identify in a multivariate situation just which vectors of observations could be considered the largest, smallest, middle 50 percent, etc.

It is clear that the practical usefulness of the technique proposed here is limited by the involvement of the solving of an integer programming problem and the lack of simplicity of least squares. However, more and more work is being done to develop algorithms which speed up computational time for solving integer programming problems. In Ref. [2] it is suggested that soon integer programs will require only between 5 and 10 times the time required by comparable linear programs. In this light it is likely that the methods here should not be considered timewise prohibitive.

BIBLIOGRAPHY

[1] Draper and Smith, *Applied Regression Analysis* (Wiley, New York, 1967).

[2] Gorry, A. and J. Shapiro, "An Adaptive Group Theoretic Algorithm for Integer Programming Problems," Management Science (Jan. 1971).

[3] Hadley, G., *Non-linear and Dynamic Programming* (Addison-Wesley, Reading, 1964).

[4] Pratt, Raiffa, and Schlaifer, *Introduction to Statistical Decision Theory* (McGraw-Hill, New York, 1965).

GENERALIZED IMPLICIT ENUMERATION USING BOUNDS ON VARIABLES FOR SOLVING LINEAR PROGRAMS WITH ZERO-ONE VARIABLES

Stanley Zionts

School of Management
State University of New York at Buffalo

ABSTRACT

In an earlier paper [11] we put forth a framework that helps to tie together a number of approaches for solving integer programming problems. We outlined there how Balas' Additive Algorithm can be explained and generalized in terms of the framework. In the present paper we review Balas' algorithm, and our earlier framework, and present an algorithm generalizing Balas' scheme. In addition, some examples are presented and future research to be done is discussed.

INTRODUCTION

The purpose of this paper is to present an algorithm for solving integer linear programming problems with zero-one variables. The algorithm—a generalization of Balas' additive algorithm—was outlined as part of a unifying approach in [11], and a relatively crude version of the algorithm was presented in [12].

We proceed by describing the problem and reviewing the method of Balas [1]. Then we review some of our earlier results and imbed Balas' framework into a more powerful algorithm. Finally, we present a numerical example, some computational experience and present some ideas for future research.

The zero-one integer linear programming problem may be formulated as follows:

$$\text{Minimize } z=\sum_{j=1}^{n} c_j x_j \tag{1}$$

$$\text{subject to: } \sum_{j=1}^{n} a_{ij}x_j\{\leqslant,=,\geqslant\}b_i \qquad i=1,\ .\ .\ .,m$$

$$x_j=0,1 \qquad j=1,\ .\ .\ .,n.$$

It is possible to rearrange any such problem so that $c_j \geqslant 0$ by making the substitution $x_j=1-\overline{x}_j$ where $\overline{x}_j$ is the complement of x_j in the constraint $x_j \leqslant 1$ where $c_j<0$. By doing this, we assure ourselves that a basic dual feasible solution, which is used as a starting point, is available. The entries c_j, a_{ij} are not required to be integer.

We now introduce the following notation:

$$x^+=\begin{Bmatrix} x & \text{if } x>0 \\ 0 & \text{if } x\leqslant 0\end{Bmatrix} \qquad x^-=\begin{Bmatrix} 0 & \text{if } x>0 \\ x & \text{if } x\leqslant 0\end{Bmatrix}.$$

A solution to a problem is the assigning of zeros and ones to the variables. A partial solution of a problem is one in which some of the variables have been specified to take on values of zero or one. Such variables are called specified, fixed, or assigned variables. Variables which are not specified in a partial solution are called *free* or *unspecified variables.* We use the convention that the leftmost element of a partial solution is the first specified variable of that solution, the second leftmost element is the second specified variable, etc. Then, a second partial solution is said to be a continuation of a first partial solution if all the elements of the first partial solution are the leftmost elements of the second partial solution. A completion of a partial solution is a continuation of a partial solution in which all free variables take on integer values. Consider the following partial solution

$$(j_1++, j_2--, j_3-, j_4-, j_5+, j_6-, j_7++, \ldots, j_p-)$$

with entries interpreted as follows: $j_k++(j_k--)$ means that x_{j_k} has been selected to be set equal to one (zero) in accordance with choice rules that are to be presented; j_k+ (j_k-) means that x_{j_k} equal to one (zero) is implied because of a partial solution of which it is a continuation. It may also mean that x_{j_k} has been set equal to one (zero) after all possible continuations of $(j_1, \ldots, j_k--,)$ $((j_1, \ldots, j_k++))$ have been (implicitly) enumerated. Thus, the partial solution $(3++, 5-, 6+, 2--)$ means that x_3 was first set equal to one by choice, then x_5 was seen to be zero (either by deduction or by having considered all consequences of x_5 being one). Similarly, x_6 must be one; then x_2 was chosen to be zero. The unsigned element j_k in a partial solution is meant to be any one of the characters j_k--, j_k++, j_k-, j_k+. (We do not make use of the choice rule j_k-- here.)
(The notation and terminology used has been evolved from that used by Balas [1], Geoffrion [4], Glover [6], and Graves and Whinston [8].)

Finally, every variable x_j shall be assumed to have a lower bound, h_j, and an upper bound, u_j, such that

$$h_j \leqslant x_j \leqslant u_j.$$

It can be seen that this is perfectly general by noting that $h_j = 0$, $u_j = 1$ designates the zero-one problem when no other information is known, and $h_j = 0$, $u_j = M$ (M arbitrarily large) designates the lower and upper bounds in a more general problem.

THE ADDITIVE ALGORITHM OF BALAS [1]

After making the earlier indicated transformation on the problem (viz. so that all $c_j \geqslant 0$) Balas sets up all the constraints as inequalities of the form:

$$\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i$$

by multiplying inequalities with the reverse inequality by -1 and using two inequalities to represent an equality. Then, if all $b_i \geqslant 0$, the optimal solution is $x_j = 0$, $j = 1, \ldots, n$. We shall present the algorithm assuming that a partial solution $(j_1, \ldots, j_p)$ is available noting that the initial partial solution has no variables assigned. Then the following procedure is used:

1a. If the current partial solution satisfies the problem constraints complete the partial solution by setting all free variables to zero, set $z^* =$ the objective function value of the solution and save the solution. No other completion of the partial solution can be better than the present one; hence no other continuations need be examined. Go to step 5.

1b. If the partial solution does not satisfy the problem constraints, go to step 2. Denote the objective function value of the current partial solution as z.

2. For *free variables* x_j such that $z+c_j<z^*$, (z^* is initialized at ∞ if no feasible solution is known initially), and such that some coefficient $a_{ij}<0$ for at least one constraint

$$b_i-\sum_{k=1}^{p} a_{ij_k}x_{j_k}<0.$$

(We shall denote the set of such variables as the set N.) Check the relationships

$$\sum_{j\epsilon N} a_{\bar{ij}} \leq b_i-\sum_{k=1}^{p} a_{ij_k}x_{j_k}\,(<0). \tag{2}$$

An intuitive interpretation of (2) is that by restricting attention to variables which can help satisfy violated constraints, we may check to see whether any constraints can never be satisfied. If any such relationship (2) is violated, there is no feasible continuation; hence, go to step 5. Otherwise go to step 3.

3a. If *all* relationships (2) are satisfied as *strict* inequalities, determine which free variable of the set N would yield the least total infeasibility (the sum of the absolute values of the amount by which all the constraints are violated) most, i.e.. choose $j_{p+1}\epsilon N$ and set $x_{j_{p+1}}=1$ such that

$$\sum_{i-1}^{m}\left(b_i-\sum_{k=1}^{p+1} a_{ij_k}x_{j_k}\right)^-$$

is maximized. The new partial solution is $(j_1, \ldots, j_p, j_{p+1}{+}{+})$. Go to step 1.

3b. If for any subset of constraints the relationship (2) holds as equalities denote by the set F all free variables x_j such that $a_{ij}<0$ for at least one constraint of the subset, and check the relationship

$$\sum_{j\epsilon F} c_j < z^*-z. \tag{3}$$

4a. If (3) is satisfied, then $x_j=1$. $j\epsilon F$ is the only possible optimal feasible continuation of the present partial solution. It may or may not be feasible, however. The next partial solution to test is therefore $(j_1, \ldots, j_p, j_{p+1}+, \ldots, j_{p+q}+)$, where $p+1, \ldots, p+q\epsilon F$. Go to step 1.

4b. If (3) is not satisfied, then there is no possible optimal feasible continuation of the present partial solution. Go to step 5.

5a. Consider the present partial solution. Find the right most element $j_k{+}{+}$ and delete all elements to the right of it. Replace it by j_k-. That is the new partial solution. Go to step 1.

5b. If there is no element $j_k{+}{+}$ in the present partial solution, the (implicit) enumeration is complete and the optimal solution (which has been saved in step 1a) has objective function value z^*. If z^* is ∞, there is no feasible solution.

Remarks: The above algorithm is the one presented by Balas [1]. Note that it yields only one optimum (it could readily be altered to yield all optima as Balas [1] points out). Note also that we have made no attempt here to implement alterations to the algorithm such as those indicated by Glover and Zionts [7] and others. In addition, a notation different from that of Balas [1] has been used here in an attempt to simplify the presentation.

THE EXTENDED GEOMETRIC DEFINITION METHOD AND ITS CONSEQUENCES FOR INTEGER VARIABLES

The Extended Geometric Definition Method is a means for computing and recomputing upper and lower bounds on (both primal and dual) variables of linear programming problems between simplex

method iterations. The bounds are then used to help determine which variables will or will not be basic in an optimal solution. We shall not pursue the Extended Geometric Definition Method further here except to present results based on it which are derived in [11]. The interested reader should consult [11] for a brief outline of the Extended Geometric Definition Method and [10] for a more complete development.

Writing each constraint i, $i=1, \ldots, m$ as

$$\sum_{j=1}^{n} a_{ij}x_j = b_i$$

we state the following theorems from [10]:

THEOREM 1: For any linear programming problem, a lower bound for a variable x_j is given by:

$$h_j = \left\{ \begin{array}{ll} \text{Max}\left\{0,\ (b_i/a_{ij}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}u_k + \sum_{k\neq j} a^{-}_{ik}h_k\right)\right\} & \text{for } a_{ij} > 0 \\ \text{Max}\left\{0,\ (b_i/a_{ij}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}h_k + \sum_{k\neq j} a^{-}_{ik}u_k\right)\right\} & \text{for } a_{ij} < 0 \end{array} \right\}.$$

PROOF: Tne proofs of all theorems are extremely straight-forward and may be found with slight variation in Zionts [10] or [11].

THEOREM 2: For any linear programming problem, an upper bound for a variable x_j is given by:

$$u_j = \left\{ \begin{array}{ll} (b_i/a_{ij}) - (1/a_{ij})\left\{\left(\sum_{k\neq j} a^{+}_{ik}h_k + \sum_{k\neq j} a^{-}_{ik}u_k\right)\right\} & \text{for } a_{ij} > 0 \\ (b_i/a_{ij}) - (1/a_{ij})\left\{\left(\sum_{k\neq j} a^{+}_{ik}u_k + \sum_{k=j} a^{-}_{ik}h_k\right)\right\} & \text{for } a_{ij} < 0 \end{array} \right\}.$$

We additionally introduce the notation that $\langle x\rangle$ the smallest integer y not less than x, i.e.,

$$\langle x\rangle = \min y \geqslant x,\ y \text{ integer}$$

and that $[x]$ is the largest integer y not greater than x, i.e.,

$$[x] = \max y \leqslant x,\ y \text{ integer.}$$

THEOREM 3: A lower bound for an integer variable x_j in a programming problem is given by

$$h_j = \left\{ \begin{array}{ll} \text{Max}\left\{0,\ \left\langle (b_i/a_{kj}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}u_k + \sum_{k\neq j} a^{-}_{ik}h_k\right)\right\rangle\right\} & \text{for } a_{ij} > 0 \\ \text{Max}\left\{0,\ \left\langle (b_i/a_{ij}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}h_k + \sum_{k\neq j} a^{-}_{ik}u_k\right)\right\rangle\right\} & \text{for } a_{ij} < 0 \end{array} \right\}.$$

THEOREM 4: An upper bound for an integer variable x_j in a programming problem is given by

$$u_j = \left\{ \begin{array}{ll} \left[(b_i/a_{ij}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}h_k + \sum_{k\neq j} a^{-}_{ik}u_k\right)\right] & \text{for } a_{ij} > 0 \\ \left[(b_i/a_{ij}) - (1/a_{ij})\left(\sum_{k\neq j} a^{+}_{ik}u_k + \sum_{k\neq j} a^{-}_{ik}h_k\right)\right] & \text{for } a_{ij} < 0 \end{array} \right\}.$$

For the zero-one integer problem, where the slacks of the constraints may not be required to be integer, Theorems 1 and 2 apply to the slack variables and Theorems 3 and 4 apply to the integer variables.

A GENERALIZATION OF BALAS' APPROACH

Consider the zero-one integer problem (1) with $c_j \geqslant 0$ with all inequality constraints in the form:

$$\sum_{j=1}^{n} a_{ij}x_j + x_{n+1} = b_i$$

and listed first for convenience and equality constraints written as they are.* We may now write the special forms of the above four theorems for this problem.

THEOREM 1′: A lower bound for a slack variable x_{n+i} is given by:

$$h_{n+1} = \text{Max}\,\{0, b_i - \sum_{k=1}^{n} a_{ik}^{+}\}.$$

PROOF: Trivial, by substitution into Theorem 1. All proofs here follow similarly and will not be given.

THEOREM 2′: An upper bound for a slack variable x_{n+1} is given by:

$$u_{n+1} - b_i = \sum_{k=1}^{n} a_{ik}^{-}.$$

THEOREM 3′: A lower bound for a zero-one integer variable is given by:

$$h_j = \begin{Bmatrix} \text{Max}\left\{0, \left\langle (b_i/a_{ij}) - (1/a_{ij}) \left(\sum_{k \neq j} a_{ik}^{+} + u_{n+i} \right) \right\rangle \right\} & \text{for } a_{ij} > 0 \\ \text{Max}\left\{0, \left\langle (b_i/a_{ij}) - (1/a_{ij}) \left(\sum_{k \neq j} a_{ik}^{-} + h_{n+i} \right) \right\rangle \right\} & \text{for } a_{ij} < 0 \end{Bmatrix}.$$

THEOREM 4′: An upper bound for zero-one integer variable is given by:

$$u_j = \begin{Bmatrix} \left[(b_i/a_{ij}) - (1/a_{ij}) \left(\sum_{k \neq j} a_{ik}^{-} + h_{n+1} \right) \right] & \text{for } a_{ij} > 0 \\ \left[(b_i/a_{ij}) - (1/a_{ij}) \left(\sum_{k \neq j} a_{ik}^{+} + u_{n+1} \right) \right] & \text{for } a_{ij} < 0 \end{Bmatrix}.$$

Where there are equality constraints $u_{n+i} = h_{n+i} = 0$. Note that the application of these theorems can be made for partial solutions by computing a value of $b_i' = b_i - \sum_{k=1}^{p} a_{ij_k}$. Then of course, only free variables are considered in the computations.

THE GENERALIZED IMPLICIT ENUMERATION ALGORITHM

By using the results of Theorems 1′ through 4′ to generate upper and lower bounds on the variables together with the Balas' structure of the implicit enumeration and by taking advantage of the

*Since each equality generates only one constraint here and is, in general, more powerful in making tests on variables, this is an improvement over Balas' algorithm.

simplification that can be obtained, a simpler and more powerful algorithm may be achieved. The resulting algorithm includes special tests developed by other authors such as Fleischmann [3] and Geoffrion [5]. Conceptually, it is convenient to pose the problem in Balas' Framework with a few alterations:

1. Equality constraints are represented as such;
2. A constraint of the form

$$\sum_{j=1}^{n} c_j x_j \leqslant z^* - \epsilon \text{ where } 0 \leqslant \epsilon \leqslant \operatorname*{Min}_j \{c_j\}$$

is added where z^* is the minimum objective function for a feasible solution found thus far. If $\epsilon = 0$ is used, then all alternate optima will be found. Initially, $z^* = M$ (where M is a sufficiently large number, e.g., Σc_j).

3. Replace steps 2, 3, and 4 of Balas' method by the following:

2. Use the results of Theorems 1′ through 4′ to generate upper and lower bounds as appropriate for each zero-one variable in every constraint.

a. If a lower bound greater than zero (but less than or equal to one) is found for some variable x_k then x_k is implied to be one in all continuations of the present partial solution. Then $k+$ augments the current partial solution. Go to step 1.

b. If a lower bound greater than one is found for some variable x_k, then there is no feasible continuation. Go to step 5.

c. If an upper bound less than one, but not less than zero is found for some variable x_k, then x_k is implied to be zero in all continuations of the present partial solution. Then $k-$ augments the current partial solution. Go to step 1.

d. If an upper bound less than zero is found for some variable x_k, then there is no feasible continuation. Go to step 5.

e. If all upper bounds are at least one and all lower bounds are at most zero, then no tighter bounds are available. Go to step 3.

3. Determine which free variable would reduce the total infeasibility (the sum of the absolute values of the amount by which all constraints are violated) most, i.e., choose j_{p+1} and set $x_{j_{p+1}} = 1$ such that

$$\sum_{i \notin E} \left| \left(b_i - \sum_{k=1}^{p+1} a_{ij_k} x_{j_k} \right)^- \right| + \sum_{i \in E} \left| \left(\sum_{k=1}^{p+1} a_{ij_k} x_{j_k} - b_i \right) \right|$$

is minimized where E is the set of equalities. The new partial solution is $(j_1, \ldots, j_p, j_{p+1}++)$. Go to step 1. (This is equivalent to Balas' choice step except that equalities are represented as such.)

4. (No step.)

Actually, the calculations in step 2 can be simplified tremendously by the elimination of tests which cannot occur, and avoiding repetitious calculations.* A flow chart for such a method is given in Figure 1. Remember that the method given in the flow chart is equivalent to that given above. An example problem will be solved by both methods.

---

*Professor Earl McCoy, University of Alabama, has suggested some of the ideas for the simplification.

EXAMPLE (Problem 1 of Balas [1]):

Minimize $z = 5x_1 + 7x_2 + 10x_3 + 3x_4 + x_5$

Subject to:

$$-x_1 + 3x_2 - 5x_3 - x_4 + 4x_5 \leq -2 \qquad x_1, \ldots, x_5 = 0, 1$$
$$2x_1 - 6x_2 + 3x_3 + 2x_4 - 2x_5 \leq 0$$
$$x_2 - 2x_3 + x_4 + x_5 \leq -1.$$

Using Balas' method, we have the following partial solutions:

1. (ϕ)
   In step 3a, add 3++.
2. (3++)
   In step 3a, add 2++.
3. (3++, 2++) $z^* = 17$; Feasible:
   In step 5a, backtrack.
4. (3++, 2−):
   In step 2, relation (2) is violated by constraint 2. Therefore backtrack (step 5a).
5. (3−)
   In step 2, relation (2) is violated by constraint 3. Therefore backtrack (step 5b). $x_3 = 1$, $x_2 = 1$ is the optimal solution.

Using the modified method we have the following sequence of solutions:

1. (ϕ)
   From constraint 3, the lower bound on x_3 is seen to be 1.
2. (3+)
   From constraint 2, the lower bound on x_2 is seen to be 1.
3. (3+, 2+) $z = 17$; Feasible.

No backtracking is possible. The problem is solved.

COMPUTATIONAL EXPERIENCE

An earlier (less efficient computationally, but equivalent otherwise) version of the algorithm has been programmed for the Graphic Controls GE−235 time sharing computer (formerly the Dartmouth facility) in the BASIC language and the results have been favorable. For two of the problems solved (two of Balas' originally published four) only one partial solution is examined before the optimal solution was known.

A few problems where comparisons were available have been solved using the computer program and the results are tabulated below:

Problem	Number of equations (m)	Number of variables (n)	Number of partial solutions solutions computed	Number of partial solutions computed using Balas' additive algorithm [1]
Balas problem 1 [1]	3	5	1	4
Balas problem 2 [1]	6	10	3	4
Balas problem 3 [1]	4	9	1	31
Balas problem 4 [1]	6	12	6	39

We have not attempted to make meaningful comparisons since we have not programmed any other methods for comparison. As indicated above, Balas' tests are a subset of ours, and hence the present algorithm will *never* require more iterations and usually less. Further, the computational requirements per iteration between this algorithm and that of Balas are about the same.

In using the algorithm for solving the problems reported as well as a number of other problems for which no comparisons were available, we became increasingly aware of the need for finding a good feasible integer solution early in the procedure. The reason for this is twofold:

1. To reduce the number of solutions that must be examined.

2. To permit a heuristic rule for stopping after a certain number of iterations, especially for larger problems where the time required for complete solution would very likely be prohibitive.

Balas' mechanism for generating a feasible solution, which we have adopted, is the following. When a variable is to be chosen to be set equal to one, choose the variable x_j which, when j++ is added to the partial solution, minimizes the measure of infeasibility defined earlier. This measure is simply the algebraic sum of the amount by which all constraints are violated. Such a mechanism gives no consideration to the cost of the chosen variable, and therefore tends only to find a feasible solution that is better than the best one found previously.

We suggest a simplex-like criterion, namely to choose the variable for which the cost per infeasibility reduction is minimum. We have tried this* for a few problems on our GE 235 BASIC program, and in every case tried the solution generated by this scheme was as good as or better than that generated by the initial scheme. For the series of 27 variable problems that we tested, it was considerably better for every problem tested.† However, the suggested scheme requires considerably greater computations than does the original scheme of Balas. It therefore seems that a reasonable way of implementing the above suggestion is to use it for generating the choice for the first q partial solutions (where q is some number to be specified) and then switch to the other choice rule. Presumably, the advantage of an early good solution would outweigh the additional time required and would be useful for continuing the procedure with Balas' original feasibility-seeking mechanism.

DISCUSSION AND FURTHER RESEARCH TO BE DONE

An improved set of tests for implicit enumeration has been developed. The method compares favorably with the additive algorithm of Balas in the following ways:

1. Equality constraints are written as they appear and are more powerful in making tests. No pairs of inequality constraints have to be used.

2. Balas' tests on the objective function are reduced to checks on one additional constraint—an objective function constraint which coincides with the objective function.

3. The present algorithm is considerably simpler than that of Balas.

4. While the present method includes all of Balas' tests, it also includes a number of tests that Balas does not have.

Unfortunately, however, it appears that any implicit enumeration scheme, regardless of how effective it is, still must examine some fraction of all possible solutions. For large problems, even the number of solutions corresponding to a tiny fraction of the total may still be enormous.

---

*Actually we deviated from the suggestion slightly. The criterion was abandoned if a feasible solution could be obtained by adding one variable to the partial solution. Also, if no infeasibility reduction is possible, then Balas' choice rule is utilized.

†These problems were provided by Professor R. Teach then at the State University of New York at Buffalo, now at Georgia Institute of Technology.

Therefore, of necessity, there is still considerable interest in methods for further accelerating the solutions. We see two promising possibilities:

1. The use of surrogate constraints as described by Balas [2], Geoffrion [5], and Glover [6]. Of the three, Geoffrion's approach appears to be most used in practice. Any comparison between the Balas additive algorithm and those methods should be carried over to the comparison between the proposed method and those methods using the proposed method.

2. An improved criterion as to which variable should be chosen to be set equal to one. We have briefly indicated one. Balas [2] has indicated another in his filter method which can be generalized as follows: For the most current surrogate constraint, choose the most efficient free variable (i.e., the one which minimizes $c_j/(b_i/a_{ij})$, $a_{ij} > 0$) to be set equal to one. There are of course other possible criteria.

In addition, there is need for extensive testing and evaluation of the schemes proposed. Our tests were made on a GE–235 time-sharing system, which unfortunately can handle only problems where $m \times n < 300$ because of core-storage limitations. Time on that system was also limiting in some instances, although that should greatly be alleviated by the simplification of the tests.

There is one final observation that we should like to make. In using our experimental code for the GE–235, we became increasingly aware of the difficulties caused by problems in which alternative variables were very similar. We shall illustrate this by the following (trivial) problem:

Minimize

$$\sum_{j=1}^{20} x_j$$

subject to

$$\sum_{j=1}^{20} x_j \geqslant 6.$$

This problem requires the examination of approximately 15,500 partial solutions before an optimum is confirmed.* The above problem can easily be made less trivial, but just as difficult to solve.

In conclusion, we have presented an algorithm into which Balas' additive algorithm can be subsumed. Although we can speak of an improvement over Balas' method with the method, it is our feeling that the real payoff of the approach will be from applying the tests to various extensions of Balas' implicit enumerative structure and from undertaking some of the research outlined above.

REFERENCES

[1] Balas, Egon, "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," Operations Research *13*, 517–546 (1965).

[2] Balas, Egon, "Discrete Programming by the Filter Method," ICC Research Report, No. 66/10, Operations Research *15*, 915–957 (1967).

[3] Fleischmann, B., "Computational Experience with the Algorithm of Balas," Operations Research *15*, 153–155 (1967).

[4] Geoffrion, A. M., "Integer Programming by Implicit Enumeration and Balas' Method," SIAM Review *9*, 178–190 (1967).

[5] Geoffrion, A. M., "An Improved Implicit Enumeration Approach for Integer Programming," Operations Research *17*, 437–454 (1969).

*However, as Fred Glover has pointed out in private communication, the use of surrogate constraints permits the solution of the above problem in very few iterations.

[6] Glover, Fred, "A Multiphase-Dual Algorithm for the Zero-One Integer Programming Problem," Operations Research *13*, 879–919 (1965).

[7] Glover, Fred and Stanley Zionts "A Note on the Additive Algorithm of Balas," Operations Research *13*, 546–549 (1965).

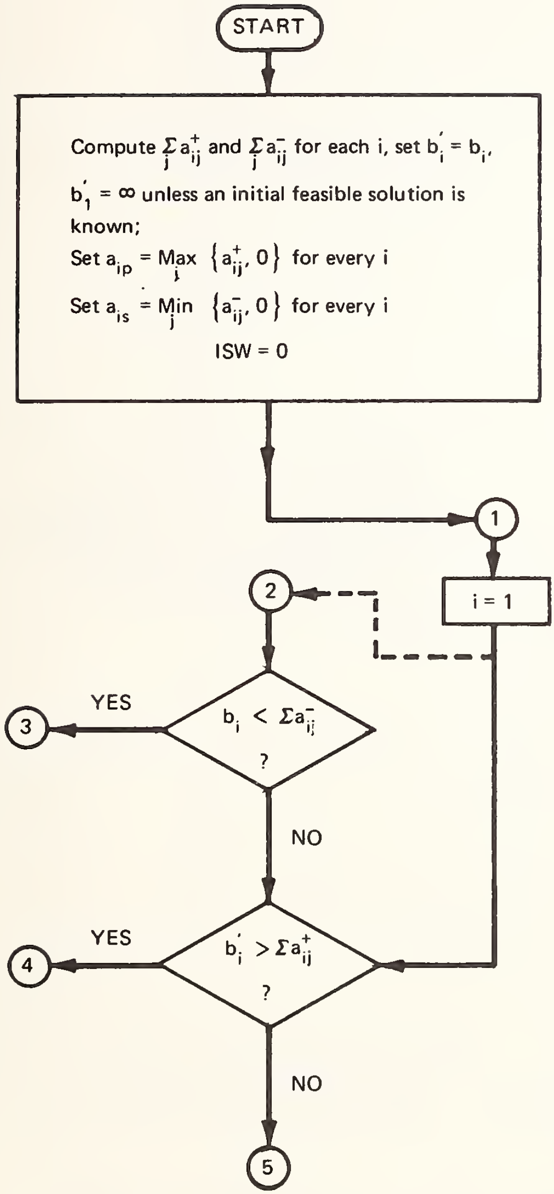
[8] Graves, G. W. and A. B. Whinston, "A New Approach to Discrete Mathematical Programming," Management Science *15*, 177–190 (1968).

[9] Rao, Ashok, "Balas' and Glover's Algorithm: A Comparison," Paper presented at the 32nd National Meeting of the Operations Research Society of America, Chicago, Illinois (Nov. 1–3, 1967).

[10] Zionts, Stanley, "Size Reduction Techniques of Linear Programming and Their Application," unpublished Ph.D. dissertation, Graduate School of Industrial Administration, Carnegie Institute of Technology, Pittsburgh, Pennsylvania (Sept. 1965).

[11] Zionts, Stanley, "Towards a Unifying Theory for Integer Linear Programming," Operations Research *17*, 359–367 (1969).

[12] Zionts, Stanley, "Implicit Enumeration Using Bounds on Variables: A Generalization of Balas' Additive Algorithm for Solving Linear Programs with Zero-One Variables," *CORSI Bulletin*. Seminar Number 1968–69, Volume 1.



It is assumed that all $c_j \geq 0$. (To relax this assumption, it is only necessary to alter one branch: see the dotted line after 1.)

Initialize and compute partial sums. The starting partial solution is the empty set. The first constraint is the objective function constraint.

Is there no feasible continuation?

For equalities, is there no feasible continuation? For inequalities, is the constraint conditionally nonbinding?

FIGURE 1. A flow chart for the generalized Balas additive method using bounds on variables [1]

GLOSSARY OF TERMS

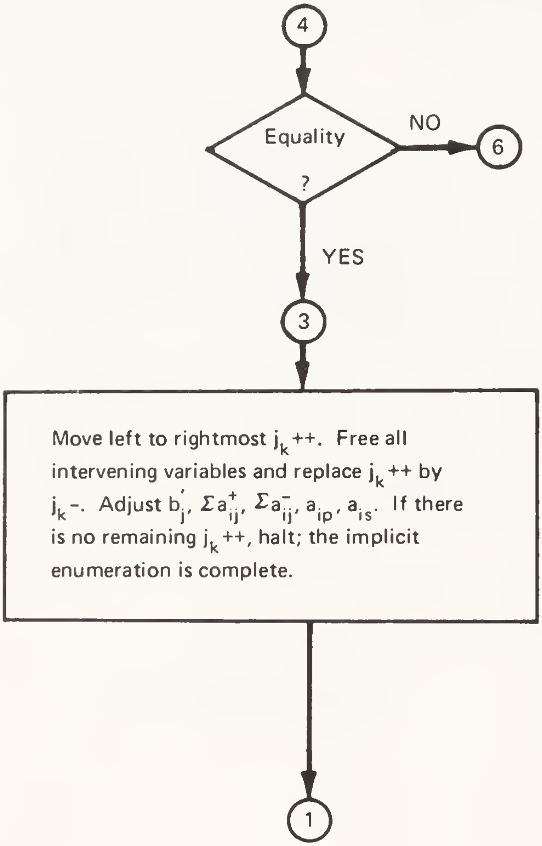
ISW — 0 if no implied variable has been detected,
1 if one or more implied variables have been detected.

$\sum a_{ij}^+$ — Sum of positive coefficients of *free variables* in row *i*.

$\sum a_{ij}^-$ — Sum of negative coefficients of *free variables* in row *i*.

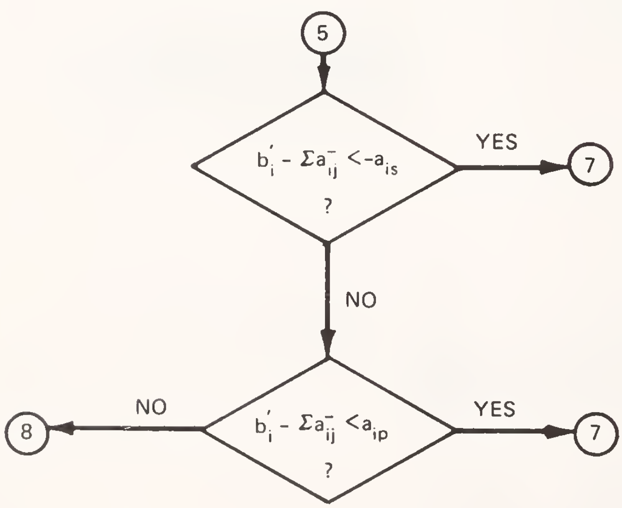
a_{ip} — The most *positive* element for a free variable in row *i*; otherwise zero.

a_{is} — The most *negative* element for a free variable in row *i*; otherwise zero.



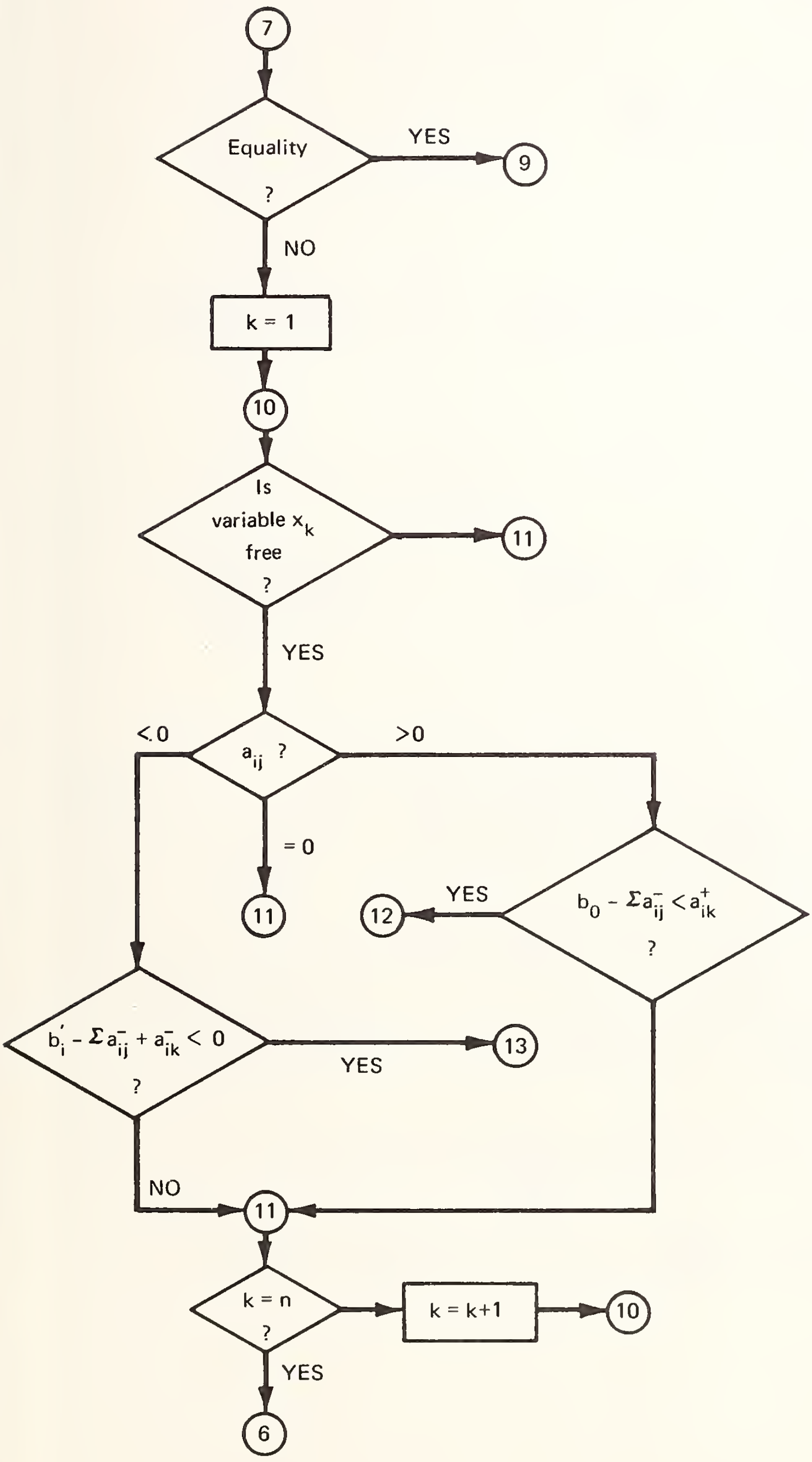
Is this constraint an equality?

Backtrack to previous selection ($x_{j_k} = 1$). Reverse the selection ($x_{j_k} = 0$) and recompute sums. If there is no previous selection step then the implicit enumeration is complete, and the stored solution is optimal.



Is the lower bound of any variable in equation i implied to be unity?

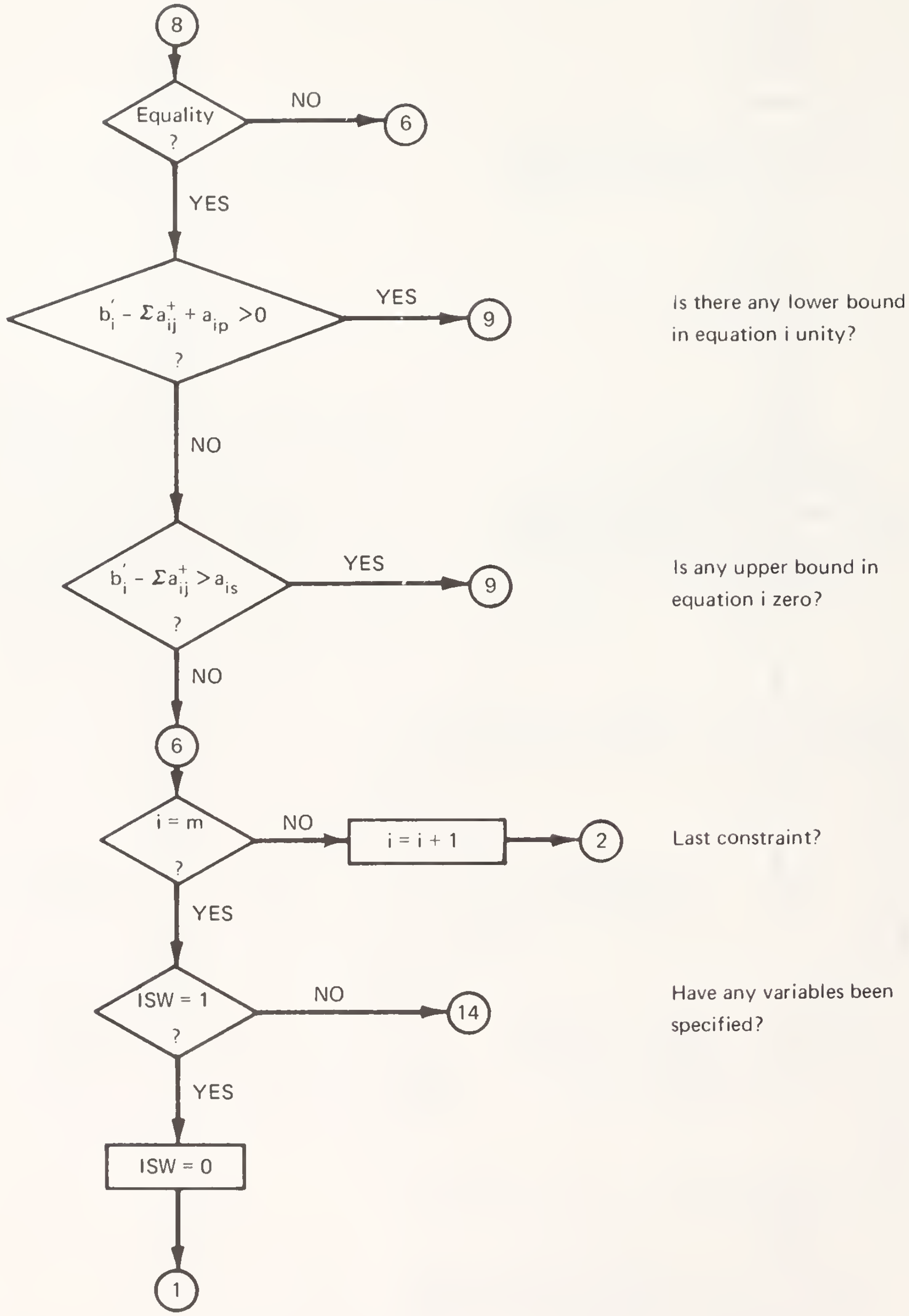
Is the upper bound of any variable in equation i implied to be zero?

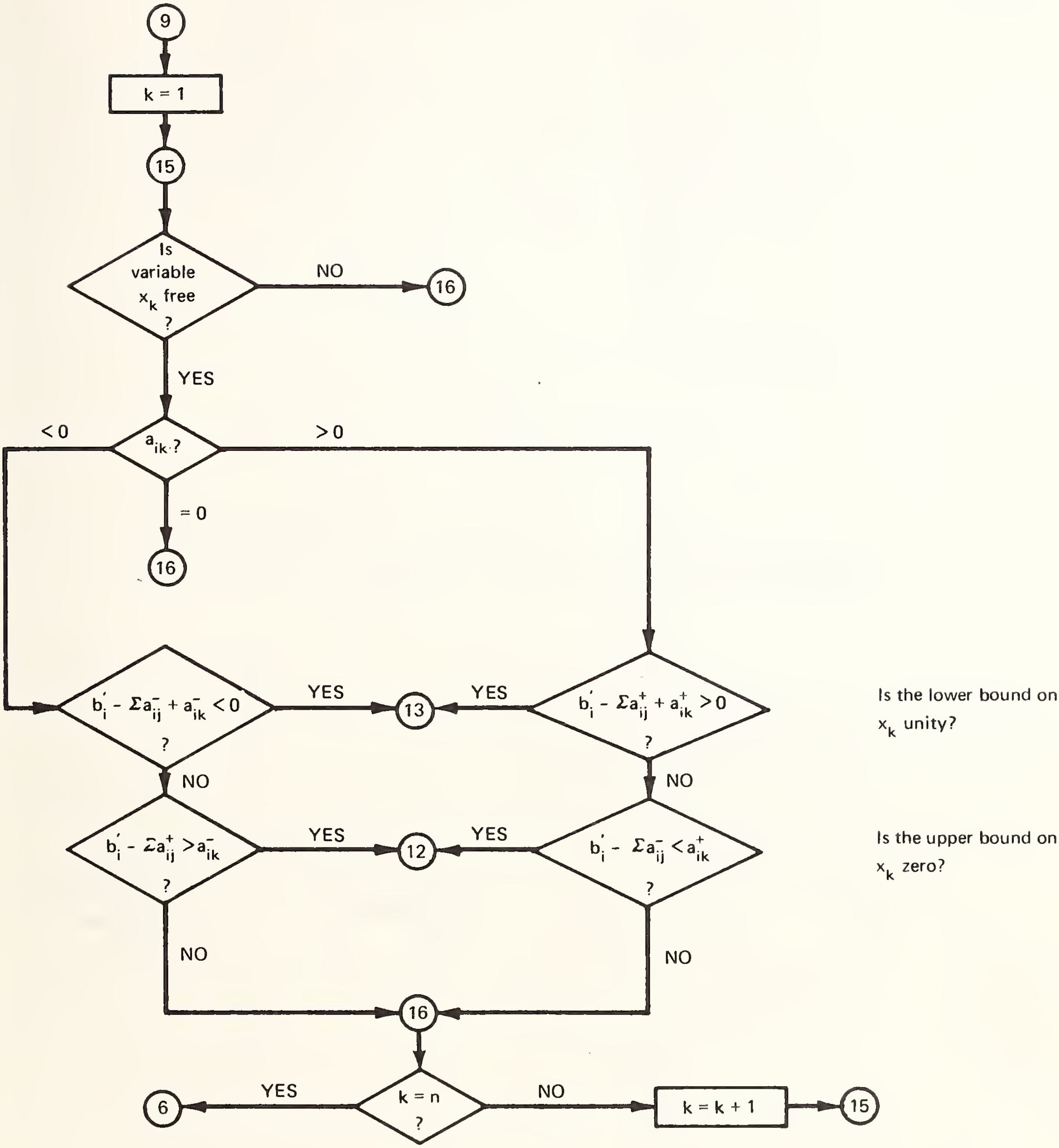


Initialize for checking

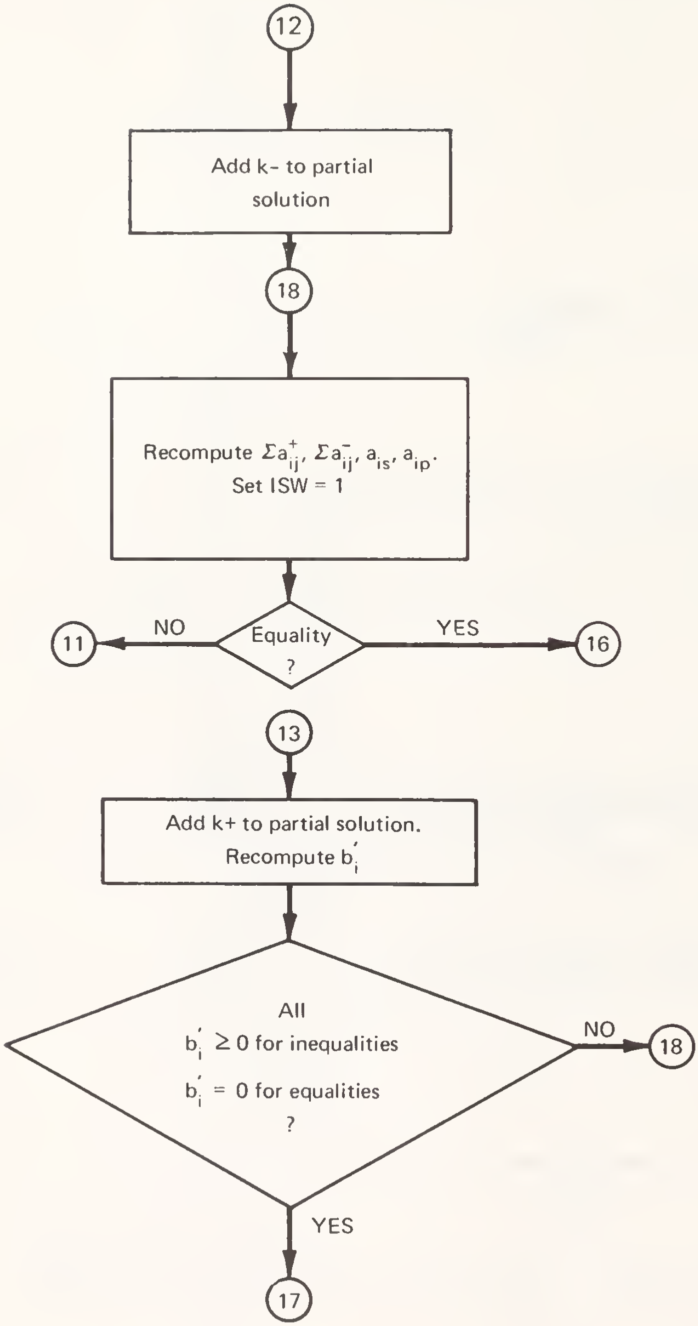
Is the upper bound on x_k zero?

Is the lower bound on x_k unity?

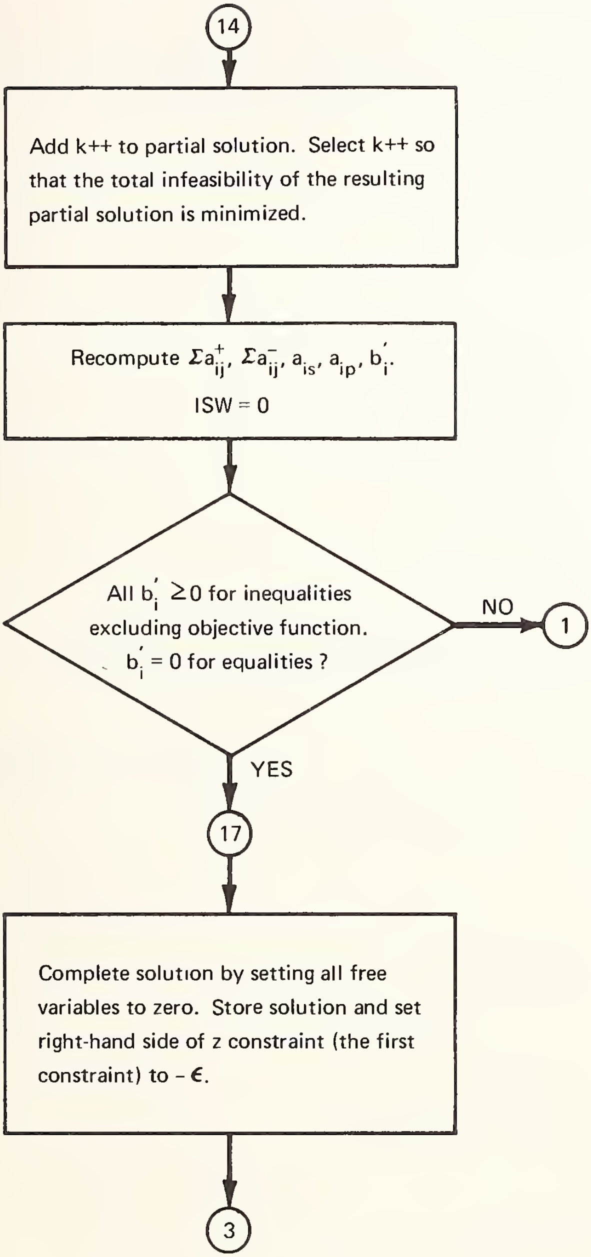

8
Equality ?
NO
6
YES
$b_i' - \Sigma a_{ij}^+ + a_{ip} > 0$?
YES
9
Is there any lower bound in equation i unity?
NO
$b_i' - \Sigma a_{ij}^+ > a_{is}$?
YES
9
Is any upper bound in equation i zero?
NO
6
i = m ?
NO
i = i + 1
2
Last constraint?
YES
ISW = 1 ?
NO
14
Have any variables been specified?
YES
ISW = 0
1



9
k = 1
15
Is variable x_k free ?
NO
16
YES
a_{ik} ?
< 0
> 0
= 0
16
$b_i' - \Sigma a_{ij}^- + a_{ik}^- < 0$?
YES
13
YES
$b_i' - \Sigma a_{ij}^+ + a_{ik}^+ > 0$?
Is the lower bound on x_k unity?
NO
NO
$b_i' - \Sigma a_{ij}^+ > a_{ik}^-$?
YES
12
YES
$b_i' - \Sigma a_{ij}^- < a_{ik}^+$?
Is the upper bound on x_k zero?
NO
NO
16
6
YES
k = n ?
NO
k = k + 1
15



Is partial solution feasible?



Choose a variable x_k to be set equal to one so that the infeasibility of the resulting partial solution is minimized.

Recompute partial sums.

Is partial solution feasible?

Feasible solution has been found. Store and tighten objective function constraint.

QUADRATIC AS PARAMETRIC LINEAR PROGRAMMING*

Robert J. Townsley†

Massey University,

New Zealand

and

Wilfred Candler

Purdue University,

Lafayette, Indiana

ABSTRACT

This paper describes an approximate solution procedure for quadratic programming problems using parametric linear programming. Limited computational experience suggests that the approximation can be expected to be "good."

PROBLEM DEFINITION

We define the quadratic programming problem:

Problem I (primal)

Maximize

(1) $$cy + \frac{1}{2} yQy$$

Subject to

(2) $$Ay \leqslant b$$

$$y \geqslant 0,$$

where Q is negative semi-definite.

Assume there exists a solution to Problem I. We denote this solution by y^*.

Problem II (dual) (Dorn [3], Moeseke [6])

Minimize

(4) $$wb,$$

Subject to

(5) $$wA \geqslant c + Qy^*,$$

(6) $$w \geqslant 0.$$

*The research reported in this paper was carried out under Project 1595 of the Purdue Agricultural Experiment Station.

†Professor of Operations Research, Massey University, New Zealand, and Professor of Agricultural Economics, Purdue University, Lafayette, Indiana, respectively.

Let w^* be a solution to Problem II.

Now, let $\bar{y}$ be any feasible solution to Problem I. If we substitute $\bar{y}$ for y^* in the statement of problem II and solve the remaining linear program we obtain $\bar{w}$. Then we have [3]:

$$\bar{w}b \geqslant c\bar{y} + \bar{y}Q\bar{y}. \tag{7}$$

Problem III

Minimize

$$wb \tag{8}$$

Subject to

$$Ay \leqslant b \tag{9}$$

$$wA - Qy \geqslant c \tag{10}$$

$$y, w \geqslant 0 \tag{11}$$

Let (w^0, y^0) be a solution to Problem III.

The following statements clearly hold.

a) The solution vector from Problem I, y^*, and the associated dual solution vector, w^*, comprise a feasible solution to Problem III.

b) Further, the solution (y^*, w^*) will be on the boundary of the feasible set for Problem III. For the case of an interior solution to Problem I we have $w^* = 0$ and y^* will still be on the boundary of Problem III, with $-Qy \geqslant c$ subject to $y \geqslant 0$.

c) Problem III is a linear programming problem, hence the solution(s) can be represented by the corner(s) of a simplex. The particular solution to Problem III, (y^*, w^*), can therefore be represented as a corner solution.

d) Corners of the simplex associated with any linear programming problem are associated with changes in the basic linear programming solution.

KUHN-TUCKER CONDITIONS

Rewrite Equations (9) and (10):

$$Ay + x = b \tag{12}$$

$$-wA + Qy + u = -c. \tag{13}$$

With respect to the quadratic programming problem: Equations (1) to (3), the Kuhn-Tucker necessary conditions [5, p. 483] can be written:

$$u, w, x, y \geqslant 0 \tag{14}$$

$$wx = yu = 0, \tag{15}$$

where x and u are the vectors of slack variables as defined in Equations (12) and (13). These conditions are necessary and sufficient provided Q is negative semi-definite.

The problem to be solved then, for Q negative semi-definite, is to find:

$$u,\ w,\ x,\ y \geqslant 0$$
$$wx = yu = 0,$$

such that Equations (12) and (13) hold. Algorithms for the solution of this problem do exist. For examples see: Boot [1], van de Panne and Whinston [9], Candler and Townsley [2], and Frank and Wolfe [4]. The method described here differs from these in that a standard linear programming code, with parametric right-hand-side routine, can be used to obtain at least approximate solutions to Problem I. The reason for consideration of an approximation method of solution where exact methods apparently exist will be made clear later.

Consider the solution to Problem III. We have:

(16)* $$w^o b \geqslant c y^o + y^o Q y^o,$$

(17)† $$w^* b = c y^* + y^* Q y^*,$$

(18)** $$w^o b \leqslant w^* b.$$

From Equation (16) we have:

$$-w^o b + c y^o \leqslant -y^o Q y^o \geqslant 0,$$

as Q is negative semi-definite.

We now define an additional constraint to be added to Problem III:

(19) $$-wb + cy = \theta,$$

where $\theta \geqslant 0$. Equations (8) to (11), plus Equation (19), now constitute Problem IV. At the optimal solution to Problem I we know that Equation (17) holds. There exists therefore some value $\theta = \theta^*$, where $\theta^* = -y^* Q y^*$ in the range $\theta \geqslant 0$. Problem IV attempts to find a value of θ such that the Kuhn-Tucker conditions for the solution to Problem I are met. Unfortunately, as we will show, even if we knew θ^*, the solution to Problem IV will not in general yield the solution to Problem I.

THE ALGORITHM

Problem IV

(20) Minimize wb.

(21) Subject to $Ay + x = b,$

(22) $$-wA + Qy + u = -c,$$

(23) $$-wb + cy = \theta, \text{ and}$$

(24) $$y,\ w,\ x,\ u \geqslant 0,$$

*See Equation (7).

†Follows directly from Kuhn-Tucker conditions, Equations (15).

**Follows from the statement of Problem III.

and where θ is varied parametrically over the range: $\theta \geqslant 0$. We consider each basic solution of this linear programming problem, with respect to the Kuhn-Tucker conditions (Equations (15)), as we vary the parameter θ over this range.*

Some empirical evidence indicates that the best approximate solution to the primal problem is attained from the above procedure when $(wx+uy)$ is a minimum. Of course, if we do obtain a solution where $wx=uy=0$ we will have satisfied the Kuhn-Tucker necessary and sufficient conditions for an optimal solution to the primal and dual problems.

We now show that the corner solutions considered by this algorithm may not include the particular solution: (y^*, w^*). It is obviously quite possible to have

$$-w^*b+cy^*=\theta^*=-w^ob+cy^o \tag{25}$$

$$w^ob < w^*b, \tag{26}$$

where (y^o, w^o) is the solution to Problem IV for $\theta=\theta^*$.

To show that this condition can hold, let y^o be a non-optimal solution to Problem I, and let w^o be the associated vector of dual variables. From Equation (7) we have

$$w^ob \geqslant cy^o+y^oQy^o,$$

therefore

$$-w^ob+cy^o \leqslant -y^oQy^o,$$

and, from Equation (25)

$$-y^oQy^o \geqslant -w^ob+cy^o=\theta^*=-w^*b+cy^*=-y^*Qy^*.$$

That is,

$$-y^oQy^o \geqslant -y^*Qy^*,$$

and

$$\frac{1}{2}y^oQy^o \leqslant \frac{1}{2}y^*Qy^*.$$

From Equations (25) and (26) we can write

$$cy^*-cy^o=-w^ob+w^*b > 0,$$

therefore

$$cy^o < cy^*.$$

Combining these last two results, we have

$$cy^*+\frac{1}{2}y^*Qy^* > cy^o+\frac{1}{2}y^oQy^o,$$

and hence, no contradiction to the hypothesis that y^o was a non-optimal solution to Problem I.

APPLICATION

The standard quadratic programming code available at this time from the Iowa State Computation

*We have already noted that the optimum solution to Problem I will be a corner solution of Problem III, and hence a corner solution of Problem IV. Consequently, we need consider only values of θ at which basis changes occur.

Center is called Zorilla [7]. This code is based on the algorithm presented by van de Panne and Whinston [9]. The algorithm solves the primal and dual problems (Problem I and II) simultaneously. Feasible solutions must satisfy Equations (9), (10), and (11) (Equations (21), (22), and (24)). In examining efficient least-cost/maximum gain poultry rations, Townsley [8] found that Zorilla was unable to find feasible solutions below 25 cents per day, even though the absolute minimum cost ration was known to cost 14.42 cents per day; and hence (due to the properties of compact convex sets) feasible solutions had to exist for costs in the range 14.42 cents to 25 cents. Because the constraint set for quadratic programming problems is linear it is extremely easy to check feasibility conditions using any standard linear programming code. This was done for our quadratic programming model, and as expected, feasible solutions for the primal and dual variables did exist for average costs in the range 14.42 cents to 25 cents. The Mathematical Programming System (MPS−360) at present in use for the solution of linear programming problems on the University IBM−360/65 computer is at least twice as accurate as the double precision arithmetic used for Zorilla. It is possible that round-off errors explain the failure of Zorilla to reach feasible solutions,* but it is also quite likely that other factors contributed to this problem.

The tentative conclusion that round-off errors were responsible for these infeasibility problems has been strengthened by a recent addition to Zorilla. This addition is a subroutine that performs any number of scaling operations on the original input data. Any reduction in the level of variability of the original matrix elements should reduce the likelihood of significant round-off errors. The completion of two scaling operations on our quadratic programming matrix enabled Zorilla to find feasible and optimal solutions in the average daily ration cost range: 25 cents to 16 cents per day. Prior to this scaling operation this cost range was infeasible as far as Zorilla was concerned. However, we have been unsuccessful in our attempts to coax Zorilla into finding feasible solutions for costs less than 16 cents per day.

In the face of these computational problems with Zorilla, the parametric linear programming algorithm already described was used to obtain approximate solutions for a number of points on the expansion path.

Table 1 presents some empirical results concerning this application of parametric linear programming.

Table 1. *Empirical Data for Six Approximate Expansion Path Rations Derived Using Parametric Linear Programming*

Ration code	Cost	θ	wx	yu
6	15.0684	43.115	0	0
7	15.0000	43.308	0.00036	0
8	14.8708	43.875	0	0.00035
9	14.8166	44.103	0	0.00008
10	14.5000	45.606	0	0
11	14.4967	45.621	0	0

From Table 1 we note that rations 6, 10, and 11 satisfy the Kuhn-Tucker conditions exactly ($wx = yu = 0$). Rations 7, 8, and 9 are approximations to their respective optima. We reiterate here that our criterion of goodness, namely magnitude of the sum ($wx + yu$), is purely empirical. Where costs equal 14.8166 cents per day, for example, ration 9 gave the highest predicted average daily gain and had the

*For this problem [7, chap. 8] the matrix Q has maximum rank 2. This sort of "near singularity" could easily give rise to round-off errors during the row and column operations of the solution process.

lowest value for the test criterion of all solutions to this particular parametric linear programming problem. We can make no attempt, however, to compare, say, rations 7 and 9 with respect to how close they come to the optimal rations for costs equalling 15.0 cents and 14.8166 cents per day, respectively.

It is possible that we could improve on the approximate solutions for rations 7, 8, and 9. For example, when cost equals 15.0 cents per day, the parametric linear programming routine indicated ration 7 as the best approximation to the expansion path (maximum gain) ration for this cost. For ration 7, $wx = 0.00036$ and the "offending" elements of w and x are w_8 and x_8. It should be possible now to employ a branch and bound technique to eliminate first one and then the other of the nonzero element pairs in w and x. However, there is no guarantee that one such step would give us the optimal solution, though of course we would reach the optimal solution in a finite number of branch and bound operations.*

The algorithm we have described entails varying the parameter θ over the range: $0 \leq \theta \leq \bar{\theta}$, where at $\theta = \bar{\theta}$ Problem IV has no feasible solution. Even though we consider only changes in the basic linear programming solution, this procedure can entail a large number of solutions. After varying θ over the entire range for two or three values of the cost function, we were able to considerably narrow the range of values for θ which were likely to be of practical significance. The values of θ corresponding to our approximate solutions are given in Table 1, and we note that as cost increases, θ decreases. Again we stress that these results are empirical and may not hold in general.

Table 2 presents a comparison between three quadratic programming solutions and their approximate linear programming counterparts. The values of the Kuhn-Tucker condition are included in Table

TABLE 2. *Comparison Between Quadratic Programming (Q.P.) and Parametric Linear Programming (L.P.) Solutions for Expansion Path Rations.*

Solution method	L.P.	Q.P.	L.P.	Q.P.	L.P.	Q.P.
OBJ (cents/day)	17.0000	17.0000	16.5694	16.5694	16.0000	16.0000
ENEG (kcal/lb)	1543	1543	1540	1540	1535	1535
PROT (percent)	13.89	13.87	13.90	13.88	13.91	13.91
INTK (lb/day)	5.0975	5.0991	5.1045	5.1062	5.1171	5.1173
GAIN (lb/day)	1.7810	1.7810	1.7805	1.7805	1.7797	1.7797
Ration ingreds (percent)						
DIST	15.95	15.79	15.93	15.78	12.68	12.67
FISH	...	...	...	...	0.98	0.98
WEAT	65.97	66.07	70.33	70.44	76.86	76.87
WHEY	8.68	8.73	4.22	4.27		
TLOW	7.00	7.00	7.00	7.00	7.00	7.00
SLYS	0.22	0.22	0.24	0.24	0.21	0.21
SMET	0.03	0.03	0.03	0.03	0.02	0.02
CACA	1.15	1.15	1.25	1.25	1.23	1.23
DCAP	...	...	...	...	0.02	0.02
Total	99.00	98.99	99.00	99.01	99.00	99.00
Kuhn-Tucker values						
wx	0.00142	...	0.00062	...	0.00004	...
yu	0	...	0	...	0	...

*This follows since the quadratic programming solution corresponds to a corner of the linear programming problem (Problem III).

2 for the linear programming solutions (these values are of course zero for the quadratic programming solutions). A perusal of Table 2 leads us to conclude that the linear programming approximations are close to their respective quadratic programming solutions.

REFERENCES

[1] Boot, John C. G., "Quadratic Programming; Algorithms—Anomalies—Applications," in Henry Theil, ed. *Studies in Mathematical and Managerial Economics.* (Rand McNally and Co., Chicago, Ill., 1964), Vol. 2.

[2] Candler, Wilfred and R. J. Townsley, "The Maximization of a Quadratic Function of Variables Subject to Linear Inequalities," Management Science *10*, 515–523 (1964).

[3] Dorn, W. S., "Duality in Quadratic Programming," Quarterly of Applied Mathematics *18*, 155–162 (1960).

[4] Frank, M. and P. Wolfe, "An Algorithm for Quadratic Programming," Nav. Res. Log. Quart. *3*, 95–110 (1956).

[5] Kuhn, H. W. and A. W. Tucker, "Non-linear Programming," in Jerzy Neyman, ed. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, Calif., 1951), pp. 481–492.

[6] Moeseke, P., "A General Duality Theorem of Convex Programming," Metro-Economica *17*, 161–170 (1965).

[7] Soults, D. J. and Janet J. Zrubek, Zorilla. Numerical Analysis-Programming Series No. 9. Ames, Iowa, Statistical Laboratory, Iowa State University (1966).

[8] Townsley, R. J., "Derivation of Optimal Livestock Rations Using Quadratic Programming," Journal of Agriculture Economics *19*, 347–354 (1968).

[9] Van de Panne, D. and A. Whinston, "The Simplex and the Dual Method for Quadratic Programming," Operational Research Quarterly *15*, 355–388 (1964).

OPTIMAL INVENTORY POLICIES IN CONTAGIOUS DEMAND MODELS*

Arunachalam Ravindran

Purdue University,
Lafayette, Indiana

ABSTRACT

In the past, contagious distributions have been successfully applied in bacteriology, entomology, and accident statistics. This paper applies the notion of contagious distributions in the inventory control of new products and seasonal or style goods, which have an underlying "true contagion" for their demands, namely, the influence of past demands on future occurrence of demands. A contagious distribution is derived by assuming a nonstationary Poisson process where the demand rate at any instant depends on the past demands prior to that instant. Using this contagious distribution, an inventory model is developed for seasonal goods and new product lines. Optimal order policies as a function of the initial stock level and the review period are derived.

1. INTRODUCTION

In the recent Proceedings of the International Symposium on Classical and Contagious Discrete Distributions [10], we observe the successful applications of the notion of contagious distributions to biological population, accident statistics, contagious diseases, and psychological data. Actually, the interest on contagious distributions dates back to 1920, when Greenwood and Yule [6] developed a very general scheme for contagious events where the occurrence of each event increases (or decreases) the probability of further events. But due to the very generality of the model, their formulas became too complex for practical applications. Neyman [9] developed and applied successfully three types of contagious distributions in entomology for distribution of larvae in experimental plots and bacteriology. Gurland [7] discusses a survey of the applications of negative binomial distribution and other contagious distributions with special reference to some medical data.

If we study the demand for new products and "style" goods (which change their style periodically), we will note that besides the constant demand for the products (which is mainly due to advertisement), a contagious demand also occurs due to customers who have used the product and recommended it to their friends or to other sources of consumer awareness. Thus, until the new product is stabilized in the market, there is a contagious demand during its transient stage. In this paper we develop an inventory system which has a contagious distribution for its system demands and discuss optimal policies for the same. Under contagion effect, given a demand has occurred from a particular area, we

*This work was partially supported by the Office of Naval Research under Contract Nonr–222(83) with the University of California, Berkeley.

expect to find more demands to come from the same area due to the influence of the past occurred demand. Thus our contagious demand model will have the underlying assumption that past demands have an influence on the occurrence of future demands.

2. APPLICATIONS OF CONTAGIOUS DEMAND

A number of examples of products can be thought of, which will exhibit a contagion demand. For example, acquiring a princess telephone in your home will tempt your neighbor to do the same. Also demands for new cereals, freeze-dried coffee, new books, automobiles (style goods), Christmas trees, and practically all consumer products in daily use, follow a contagious law. The notion of contagion can even be extended to nonconsumer products like reprints of research reports which exhibit a contagion pattern. Also, the sequence of published research papers in a particular field will tend to follow a contagious law. Another classical application is the efficient use of hospital beds for patients with contagious diseases.

3. A MULTIPERIOD CONTAGIOUS DEMAND MODEL

In this section, a discrete contagious demand model will be developed for new product lines and style goods, under the assumption that every past demand influences further occurrence of demands. We do this by assuming a nonhomogeneous Poisson process whose demand rate $\lambda(t)$ at any instant t given r previous demands before t, is given by $\lambda(t)=\lambda+\alpha(t)r$ where $\lambda(>0)$ is the constant demand rate and $\alpha(t)$ is unit contagious demand rate. As one can normally expect, the contagion rate $\alpha(t)$ may be high during the initial phase of a new product and then may follow a decay law such that its effect vanishes after the product is stabilized in market. Thus $\alpha(t)$ is in fact a decreasing function of time. But to simplify our analysis we make some assumptions about the variation of $\alpha(t)$.

The seasonal or style goods change annually or semiannually. Hence an approximation may be made by assuming a constant (average) contagion rate throughout the season. This may be valid if the season is small or if the contagion rate is changing very slowly.

The nonseasonal goods like new cereals will be in the market for a considerable length of time. So in this case the contagion rate is approximated as a step function with breaks at regular time intervals identified as the review period, T. Choice of T is made institutionally depending upon how rapidly the contagion rate is changing. Thus the contagion rate is assumed to be constant in each review period. Of course, the contagion rate will be less in each succeeding period and its new value at the end of every review period will be estimated by the knowledge of past realizations of demand.

4. A CONTAGIOUS PROBABILITY DISTRIBUTION FOR DEMANDS

In the previous section, it was indicated that only a single period inventory model will be used for seasonal goods while a multiperiod inventory model will be developed for new product lines (nonseasonal goods). Let us first confine ourselves to the first review period when a new product has just been introduced. Let T denote the length of the review period. Consider an instant of time $t\epsilon(0, T)$. The total demand rate at any instant t, given r previous demands in $(0, t)$ is given by

$$\lambda(t)=\lambda+\alpha r \qquad \text{for } r=0, 1, 2, \ldots, \tag{4.1}$$

where $\alpha(>0)$ is the unit contagious demand rate assumed to be constant. Assuming a nonhomogeneous

Poisson process where the demand rate at any instant, t, is given by (4.1), we can derive the contagious probability distribution as

$$P_n(t) = \text{Probability of } n \text{ demands in } (0, t) = [\lambda(\lambda+\alpha) \ldots\ldots\ldots\ldots (\lambda+\alpha(n-1))/$$

$$\alpha^n n!] \exp(-\lambda t)[1-\exp(-\alpha t)]^n \qquad \text{for } n=0, 1, 2 \ldots.$$

Using the well known Gamma notation:

$$P_n(t) = [\Gamma(\rho+n)/\Gamma(\rho)\Gamma(n+1)][\exp(-\alpha t)]^\rho [1-\exp(-\alpha t)]^n, \tag{4.2}$$

where $\rho = \lambda/\alpha =$ a positive constant, but need not be an integer. Equation (4.2) may also be interpreted in the same sense as a negative binomial distribution by suppressing the time parameter, t. This negative binomial characteristic of the contagious distribution is true for many other contagious distributions in literature. The introduction of α and t distinguishes our contagious distribution from classical negative binomial distributions.

Since the contagious distribution reduces to a negative binomial distribution given t and ρ, the tables of the latter may be used to find the probabilities of the contagious distribution. Tables of negative binomial distribution are available in [15]. Also, Taylor [14] shows an important mathematical equality in his paper that the infinite sum of negative binomial terms can be expressed as a finite sum of positive binomial terms. Hence, the latter's table can be used to find the former.

It can be verified that the generating function of $P_n(t)$ is:

$$G_t(z) = (p'/1-q'z)^\rho \qquad \text{where } p' = \exp(-\alpha t),\ q' = 1-p'. \tag{4.3}$$

From (4.3), we gan get the mean and variance of the contagious distribution as

$$m(t) = \text{Expectation of } P_n(t) = \rho q'/p', \tag{4.4}$$

and

$$V(t) = \text{Variance of } P_n(t) = pq'/(p')^2. \tag{4.5}$$

For a multiperiod inventory system, we need to know the demand distribution in second and succeeding periods. It turns out that the distribution varies from one period to another. This is mainly because of the effect of past demands and the decreasing nature of contagion rate. The demand distribution for review period i is derived by assuming the number of demands in the first $(i-1)$ periods as N_i and the current estimated value of the contagion rate as α_i. By our assumption $\alpha_i \leqq \alpha_{i-1} \leqq \ldots \leqq \alpha_1$.

Once again, the beginning of the ith period will be denoted as time zero. Then, the rate at which demands occur at any instant t, given r demands during $(0, t)$ will be given by

$$\lambda(t) = \lambda + \alpha_i N_i + \alpha_i r. \tag{4.6}$$

The term $(\lambda + \alpha_i N_i)$ can be replaced by a constant rate λ_i. Comparing (4.6) with (4.1), we can immediately

get the demand distribution in ith period given N_i as

(4.7) $P_n^{(i)}(t) = [\Gamma(\rho_i+n)/\Gamma(\rho_i)\Gamma(n+1)]\,[\exp(-\alpha_i t)]^{\rho_i}[1\text{-}\exp(-\alpha_i t)]^n$ for $n=0, 1, 2, \ldots,$

where $\rho_i = \lambda_i/\alpha_i = (\lambda + N_i\alpha_i)/\alpha_i$.

(4.7) implies that the probability distribution for demands in ith review period depends on the number of demands that occurred in the previous periods.

Since it is postulated that the contagion rates $\alpha_1, \alpha_2, \ldots, \alpha_i \ldots$ are decreasing, it will be of interest to know the limiting form of (4.7) as the contagion rate vanishes. Using the generating function of (4.7), it can be shown that $P_n^{(i)}(t)$ tends to a Poisson distribution as α_i tends to zero.

5. A SINGLE PERIOD INVENTORY MODEL

We discussed earlier that for the case of seasonal goods a single period inventory model is used with a constant contagion rate throughout the season. On the other hand, in the case of (nonseasonal) new product lines, a multi-period inventory model with a periodic review policy will be developed where the contagion rate decreases in each succeeding period but remains constant during any review period. This calls for new estimated values of the parameters of the demand distribution after every review period. Because of this feature, the multi-period problem is treated as successive single period problems. At the beginning of each review period, using the past realization of demand, the new value of the contagion rate is estimated, which along with the knowledge of past demands, gives the contagious demand distribution for that period. To facilitate looking at the system more often initially, one may choose the initial review periods of much smaller length compared to that of later periods.

Now, let us turn our attention to the inventory system under contagion demand where the probability demand distribution is given by

(5.1) $P_n(T) = [\Gamma(\rho+n)/\Gamma(\rho)\Gamma(n+1)][\exp(-\alpha T]^{\rho}[1-\exp(-\alpha T)]^n$ for $n=0, 1, 2, \ldots.$

It is assumed that for the current review period, T, the values of the parameters of the contagious distribution given by (5.1) are known. The following conventional assumptions will be made about the inventory system:

(i) No disposal of goods at the end of any review period.

(ii) Goods are reordered at the beginning of the review period and there is instantaneous delivery.

(iii) Unsatisfied demands may either be backlogged or sales lost.

Let x denote the initial inventory level before reordering and y denote the inventory level after reordering. Hence $(y-x)$ is the amount reordered. Our aim is to find an optimal amount to be reordered which will maximize the expected net revenue in that period. We shall give an acceptable parametric form for the net revenue function which will facilitate studying the optimal order policies as a function of the various cost parameters. The conventional costs like the acquisition cost, inventory holding cost and back order cost will be considered in computing the expected net revenue. Because of the earlier assumptions, only the following two cases have to be considered in computing the various cost components:

CASE (i): $y \geqq 0$, $y \geqq x$ (x may be positive, negative, or zero).

CASE (ii): $y < 0$, $y \geqq x$.

A typical inventory cycle under contagion demand is illustrated in Figures 1 and 2 for Cases (i) and (ii), respectively. A negative inventory level indicates excess demand over supply and orders

backlogged.

The expected gross revenue corresponds to the goods sold in the current period plus the last period's backorders filled at the beginning of the period. Denoting the unit selling price by r, we get the following expressions for the two cases:

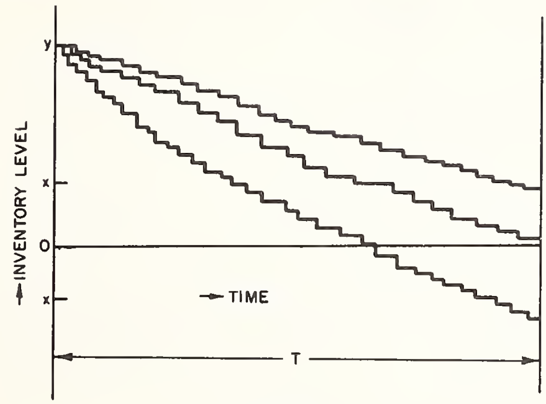


FIGURE 1

CASE (i): (Refer to Figure 1).

The expected gross revenue is

$$-r \min(x, 0) + rm(T) - r \sum_{n=y+1}^{\infty} (n-y)P_n(T),$$

where $P_n(T)$ is the contagious probability distribution of demand given by (5.1), and $m(T)$ is its expected value for a given review period, T.

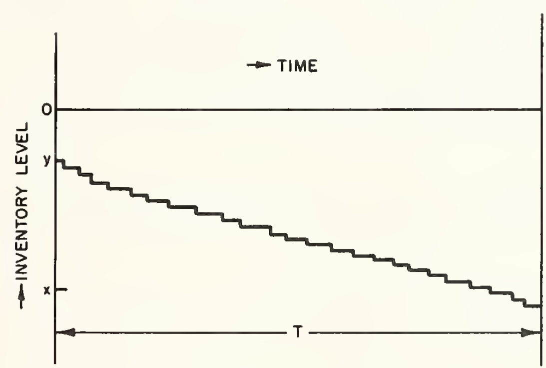


FIGURE 2

CASE (ii): (Refer to Figure 2).

The expected gross revenue is $r(y-x)$. (Note in this case both x and y are negative.) The acquisition cost is given by a setup cost (K), independent of amount reordered and a unit cost of ordering (c). This cost will be same for both Cases (i) and (ii) and is given by

$$K\delta(y-x) + c(y-x), \text{ where } \delta(y-x) = 1 \quad \text{if } y > x$$

$$= 0 \quad \text{otherwise.}$$

The holding cost will be proportional to the instantaneous inventory level and how long each unit is held in inventory. Let h denote the holding cost/ unit time/ unit held in inventory. Then for Case (i), a conventional calculation of holding costs gives the value:

$$h \sum_{n=0}^{y} (y-n) \int_0^T P_n(t)dt.$$

By using (5.1) and the appendix, the quantity

$$\int_0^T P_n(t)dt = I_q(n+1, \rho)/\alpha(\rho+n), \tag{5.2}$$

where $q=1-\exp(-\alpha T)$ and $I_q(n+1, \rho)=$ Incomplete Beta Function Ratio with parameters q, $n+1$ and ρ.

By using (5.2), the expected inventory holding cost is

$$h \sum_{n=0}^{y} (y-n) I_q(n+1, \rho)/\alpha(\rho+n) \qquad \text{for Case (i).}$$

For Case (ii), the holding cost is zero since no inventory is carried. (Tables of incomplete Beta function ratios are available in Pearson [11].)

We consider two types of backorder costs: one depending on the number of units backordered and another depending on the length of time an order remains unfilled. The parameters for the above costs are denoted by p and p^* where p is the unit backorder cost and p^* is the unit backorder cost per unit time. A completely conventional calculation of backorder costs gives the value:

$$p \sum_{n=y+1}^{\infty} (n-y) P_n(T) + p^* \sum_{n=y+1}^{\infty} (n-y) I_q(n+1, \rho)/\alpha(\rho+n) \qquad \text{for Case (i),}$$

and

$$(p+p^*/\alpha) m(T) - p^* \rho T - (p+p^* T) y \qquad \text{for Case (ii).}$$

The expected net revenue function for a given review period, T, with order level, y, is:

$\pi(y, T) =$ Expected gross revenue $-$ Acquisition cost $-$ expected holding cost $-$ expected backorder cost.

It is possible to write a general expression for $\pi(y, T)$ combining Cases (i) and (ii) as follows:

$$\pi(y, T) = rm(T) - r \min(x, 0) + cx - K\delta(y-x) - G(y, T), \tag{5.3}$$

where

$$(5.4)\quad G(y, T) = cy + (p+r) \sum_{n=0}^{\infty} [n - \min(y, n)] P_n(T)$$
$$+ h \sum_{n=0}^{\infty} [\max(y, n) - n] I_q(n+1, \rho)/\alpha(\rho+n)$$
$$+ p^* \sum_{n=0}^{\infty} [n - \min(y, n)] I_q(n+1, \rho)/\alpha(\rho+n).$$

The expressions for $\pi(y, T)$ for Case (i) may be obtained easily by putting $y \geqq 0$, $y \geqq x$ in (5.3) and (5.4). Similarly, putting $y < 0$, $x \leqq 0$, $y \geqq x$ gives $\pi(y, T)$ for Case (ii).

6. OPTIMAL ORDER POLICIES

We now turn to the main section of the paper, the determination of optimal policies. We shall first find the best order level y for a given T which will maximize $\pi(y, T)$ and then discuss how the order policies vary as the review period changes. From (5.3) it is clear that maximizing $\pi(y, T)$ over y for a given T is equivalent to minimizing the cost function $K\delta(y-x) + G(y, T)$. Since $K\delta(y-x)$ is a step function of y, the main interest is to examine the propertie t. he cost function $G(y, T)$ for all integer values of y.

PROPOSITION 6.1: The cost function $G(y, T)$ is strictly pointwise convex in $y \epsilon \{0, 1, 2, \ldots\}$ for given $T \geqq 0$.

PROOF: Putting $y \geqq 0$ in (5.4) we get

$$(6.1)\quad G(y, T) = cy + (p+r) \sum_{n=y+1}^{\infty} (n-y) P_n(T)$$
$$+ h \sum_{n=0}^{y} (y-n) I_q(n+1, \rho)/\alpha(\rho+n)$$
$$+ p^* \sum_{n=y+1}^{\infty} (n-y) I_q(n+1, \rho)/\alpha(\rho+n).$$

The first difference of $G(y, T)$ is:

$$(6.2)\quad \Delta G(y, T) = G(y+1, T) - G(y, T) = -(p+r-c)$$
$$+ (p+r) \sum_{n=0}^{y} P_n(T) + h \sum_{n=0}^{y} I_q(n+1, \rho)/\alpha(\rho+n)$$
$$- p^* \sum_{n=y+1}^{\infty} I_q(n+1, \rho)/\alpha(\rho+n).$$

The second difference of $G(y, T)$ is:

$$\Delta^2 G(y, T) = \Delta G(y+1, T) - \Delta G(y, T) = (p+r) P_{y+1}(T)$$
$$+ (h+p^*) I_q(y+2, \rho)/\alpha(\rho+y+1) > 0$$

since $P_{y+1}(T)$ and $I_q(y+2, \rho)$ are positive for all $y \epsilon (0, 1, 2, \ldots)$. Hence the proposition holds.

To investigate $G(y, T)$ for negative integer values of y, put $y \leqq 0$ in (5.4) to get:

(6.3) $G(y, T) = (p + r + p^*/\alpha)m(T) - p^*\rho T - (p + p^*T + r - c)y \qquad \text{for } y\epsilon\{0, -1, -2, \ldots\}.$

Using (6.1) and (6.3), it is clear that $G(y, T)$ is continuous at $y = 0$.

Critical Order Level

Let $y_0(T)$ denote the critical order level for a given T that minimizes $G(y, T)$. From (6.3), it is clear that $G(y, T)$ is linear in y with negative slope for $y\epsilon\{-1, -2, \ldots\}$. Hence the absolute minimum of $G(y, T)$ cannot occur in that region. Hence, it must occur for some $y\epsilon\{0, 1, 2, \ldots\}$. Since $G(y, T)$ is strictly pointwise convex in this region (Proposition 6.1), the point $y_0(T)$ that minimizes $G(y, T)$ is unique. Note that, if the slope of $G(y, T)$ is positive at $y=0$, then the minimum of $G(y, T)$ occurs at $y = 0$. If, on the other hand, the slope is nonpositive, then the minimum occurs for some positive integer value of y. Hence it is necessary to investigate the sign of $\Delta G(y, T)$ at the origin. Let $a(T) = \Delta G(0, T)$. Putting $y = 0$ in (6.2), we get after some simplification:

(6.4) $$a(T) = -(p + r + p^*T - c) + (h + p^*)/\lambda + P_0(T)[(p + r) - (h + p^*)/\lambda].$$

If $a(T) > 0$ for a given T, then $y_0(T) = 0$. On the other hand, if $a(T) \leqq 0$ then $y_0(T) > 0$ (integer). From (6.4), as T tends to zero, $a(T)$ tends to a positive value; while as $T \to \infty$, $a(T)$ tends asymtotically to $-(p + r - c) + (h + p^*)/\lambda - p^*T$ which is negative and remains so for sufficiently large values of T.

PROPOSITION 6.2: There exists a unique and finite $T_0 > 0$, such that $a(T_0) = 0$, $a(T) > 0$ for $T < T_0$ and $a(T) < 0$ for $T > T_0$.

PROOF: Differentiating a (T) with respect to T:

$$a'(T) = [(h + p^*)/\lambda - (p + r)]\lambda P_0(T) - p^*$$

and

$$a''(T) = -[(h + p^*)/\lambda - (p + r)]\lambda^2 P_0(T).$$

When $(h + p^*)/\lambda \leqq (p + r)$, $a(T)$ is a decreasing function of T and the proposition follows immediately. When $(h + p^*)/\lambda > (p + r)$, $a(T)$ is a strictly concave function of T. Hence, $a'(T)$ becomes negative after some T. Hence the proposition is true.

From Proposition (6.2), it follows that

(6.5) $$y_0(T) = 0 \quad \text{for all} \quad 0 \leqq T < T_0$$
$$= \text{positive integer for } T \geqq T_0.$$

For $T \geqq T_0$, the critical order level $y_0(T)$ is sought such that

(6.6) $$\Delta G(y_0(T) - 1, T) \leqq 0$$

$$\Delta G(y_0(T), T) > 0.$$

A typical illustration of $G(y, T)$ with respect to y, for a given T, is given in Figure 3.

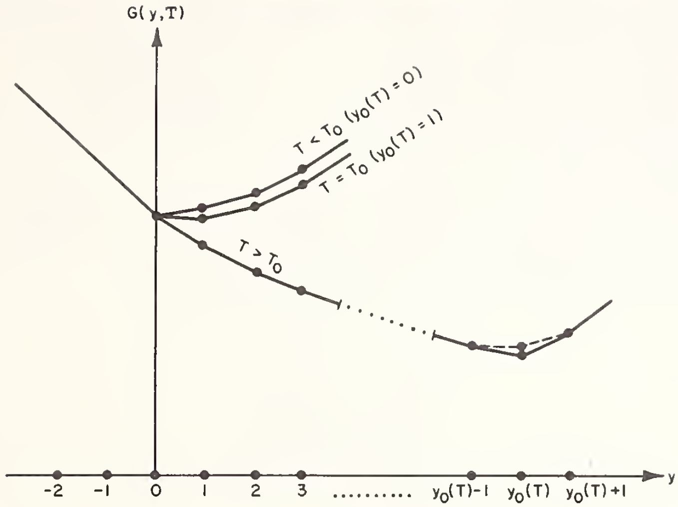


FIGURE 3

Optimal Order Level

Given an initial order level x, if we order up to level $y_0(T)$, then the total cost of ordering and operating at level $y_0(T)$ is $K+G(y_0(T), T)$. Instead if we did not order and operate at level x, then the cost is $G(x, T)$. Hence, the optimal order policy will be to order up to $y_0(T)$ only if

$$G(x, T) > K+G(y_0(T), T).$$

Otherwise no stock is ordered. Hence, the optimal order level denoted by $y^*(T)$ will be

$$y^*(T) = \begin{cases} y_0(T) & \text{if } G(x, T) > K+G(y_0(T), T) \\ x & \text{otherwise.} \end{cases}$$

Such an order policy is known as an (s, S) type order policy in inventory literature, where $S = y_0(T)$ and s is the break-even value of x such that $G(s, T) = K + G(y_0(T), T)$. The (s, S) policy is stated as "order up to S if and only if the inventory level is less than s".

To find $y_0(T)$, the expression (6.6) has to be solved. Since we have shown that $y_0(T)$ is unique for a given T, we may apply any one dimensional search algorithms, like the bisection method or Fibonacci search, to find $y_0(T)$ which in turn gives the optimal order policy. For a detailed explanation of the bisection algorithm, refer to [13].

7. AN EXAMPLE

In this section, we shall present a numerical example showing how to calculate the critical order level $S = y_0(T)$, and the break-even value s for the (s, S) order policy.

Consider the following data for a specific case:

r (selling price)	= \$1400 per unit
c (ordering cost)	= \$700 per unit
h (inventory holding cost)	= \$4 per unit per week
p (constant backorder cost)	= \$100 per unit

p^* (time dependent backorder cost) $= \$6$ per unit per week
λ (constant demand rate) $= 2$ per week
α (contagious demand rate) $= 1$ per week
K (set up cost) $= \$500$
$\rho = \lambda/\alpha = 2$

We shall carry out the analysis for three different review periods; $T = 5$ days, $T = 10$ days, and $T = 15$ days.

Using (6.6), the critical order level $S = y_0(T)$ is sought such that

$$\Delta G(S-1, T) \leqq 0 \tag{7.1}$$

$$\Delta G(S, T) > 0,$$

where

$$\Delta G(N, T) = -(p+r-c) + (p+r)\sum_{n=0}^{N} P_n(T) \tag{7.2}$$

$$+ h \sum_{n=0}^{N} I_q(n+1, \rho)/\alpha(\rho+n)$$

$$- p^* \sum_{n=N+1}^{\infty} I_q(n+1, \rho)/\alpha(\rho+n) \qquad \text{for all } N = 0, 1, 2, \ldots.$$

It has been shown in section 4 that $P_n(T)$ has a negative binomial distribution for a given T. Thus, using the tables of negative binomial distribution given by Williamson and Bretherton [15], the values of $P_n(T)$ can be obtained for every n for the three review periods. Similarly, the values of the incomplete beta function ratios, $I_q(n+1, \rho)$, can be read off directly from Pearson's table [11]. It may be pointed out here that the value of $I_q(n+1, \rho)$ decreases as n increases and tends to zero for some large value of $n = M$. Thus, in Equation (7.2), the infinite summation for the last term is carried out only up to $n = M$. A small computer program was written to evaluate Equation (7.2) for the three different review periods and the results are given below:

CASE (i): $T = 5$ days

$N =$	0	1	2
$\Delta G(N, T) =$	-425	-49	$+232$

Equation (7.1) is satisfied by $N = 2$. Hence, the critical order level $S = 2$.

CASE (ii): $T = 10$ days

$N =$	0	1	2	3	4	5	6
$\Delta G(N, T) =$	-717	-583	-432	-279	-134	-1.2	$+116$

Critical order level $S = 6$.

CASE (iii): $T = 15$ days

$N =$	0	1	2	3	4	5	6
$\Delta G(N, T) =$	-786	-746	-694	-634	-568	-498	-426

$N =$	7	8	9	10	11	12	13
$\Delta G(N, T) =$	-355	-284	-214	-148	-84	-24	$+33$

Critical order level $S = 13$.

It may be observed in the above example that the critical order level increases as the review period increases. This is true in general and the following propositions are proved in Reference [13]:

(i) The critical order level is a nondecreasing step function of the review period.

(ii) The critical order level is bounded above as the review period tends to infinity.

After having determined the order level S, we have to determine the breakeven value s, such that $G(s, T) = K + G(S, T)$ for the optimal (s, S) order policy. A small computer program was written to compute $G(N, T)$, where

$$(7.3)\quad G(N, T) = CN + (p + r) \sum_{n=N+1}^{\infty} (n - N)P_n(T) + h \sum_{n=0}^{N} (N - n)I_q(n + 1, \rho)/\alpha(\rho + n) + p^* \sum_{n=N+1}^{\infty} (n - N)I_q(n + 1, \rho)/\alpha(\rho + n) \qquad \text{for } N = 0, 1, 2, \ldots.$$

For the setup cost $K = 500$, the break-even values of s are given below for the three review periods:

$$s = \begin{cases} 0 \text{ for } T = 5 \text{ days} \\ 3 \text{ for } T = 10 \text{ days} \\ 9 \text{ for } T = 15 \text{ days.} \end{cases}$$

Thus, for example, when the review period is 10 days, the optimal policy is to order up to six units only if the inventory level is less than three units.

8. CONCLUDING REMARKS

The results obtained so far had the assumption that the inventory system follows a backordering policy. Instead, if a lost-sales policy were followed, then the backorder cost parameters p and p^* can be set to zero and all the results of the previous sections will follow identically. Instead, if we

want to associate a good will loss with each sales lost, then p can be made positive and the same results hold.

Also we had the assumption that the review periods for (nonseasonal) new product lines are chosen institutionally. But for seasonal goods we have only a single period to consider. Then we may ask the question, instead of choosing that review period institutionally, is it possible to find an optimal review period which will maximize the total net revenue. This has been successfully carried out in the author's original report [13] where the net revenue function was first maximized with respect to y for a given T and then substituting the optimal value of y in the revenue function, the best review period, T, was determined. The statistical estimation methods for the contagious distribution have also been discussed in [13]. These results will be published shortly.

9. ACKNOWLEDGMENTS

The author wishes to express his deepest appreciation to Professor R. W. Shephard (University of California, Berkeley) for introducing him to this topic and guiding the progress of this research from its inception.

Appendix

BETA FUNCTIONS

By definition, a Beta Function of two parameters m, n is given by

$$B(m, n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx$$

$$= \Gamma(m)\Gamma(n)/\Gamma(n+m), \qquad \text{where } m > 0,\ n > 0.$$

An incomplete Beta function is defined as

$$B_x(m, n) = \int_0^x y^{m-1}(1-y)^{n-1}\,dy, \qquad \text{where } x < 1$$

The ratio of incomplete Beta function to the (complete) Beta function is known as the incomplete Beta function ratio and is denoted by

$$I_x(m, n) = \frac{B_x(m, n)}{B(m, n)}.$$

Note that $I_1(m, n) = 1$. Also,

$$I_x(m, n) = 1 - I_y(m, n) \qquad \text{where } x + y = 1$$

Tables of incomplete Beta functions and ratios are available in Pearson [11]. It can be verified when m is an integer that

$$I_x(m, n) = 1 - (1-x)^n \sum_{i=0}^{m-1} \binom{n-1+i}{i} x^{i-1}.$$

The advantage of this equation is that of computational feasibility. The above series is made up of en-

tirely positive terms and hence can be summed quite accurately, even for fairly large values of parameters m and n.

BIBLIOGRAPHY

[1] Cox, D. R. and H. D. Miller, *The Theory of Stochastic Processes* (John Wiley and Sons, Inc., New York, 1965).

[2] Cramer, H., *Mathematical Methods of Statistics* (Princeton University Press, Princeton, New Jersey, 1946).

[3] Eggenberger, F. and G. Polya, "Uber die Statistik ver Ketteter Vorgange," Zeits-chrift fur Angewandte Mathematik und Mechanik **1**, 279–289 (1923).

[4] Feller, W., *An Introduction to Probability Theory and Applications* (John Wiley and Sons, Inc., New York, 1957), vol. 1.

[5] Feller, W., "On a General Class of Contagious Distributions," Annals of Mathematical Statistics **14**, 389–400 (1943).

[6] Greenwood, M. and G. U. Yule, "An Enquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Disease or Repeated Accidents," Journal of the Royal Statistical Society **83**, 255–279 (1920).

[7] Gurland, J., "Some Applications of Negative Binomial and Other Contagious Distributions," American Journal of Public Health **49**, 1388–1399 (1959).

[8] Hadley, G. and T. M. Whitin, *Analysis of Inventory Systems* (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[9] Neyman, J., "On a New Class of Contagious Distributions," Annals of Mathematical Statistics **10**, 35–57 (1939).

[10] Patil, G. P., "Classical and Contagious Distributions," Proceedings of the International Symposium held at McGill University, Montreal, Canada, Statistical Publishing Society, Calcutta, India (1965).

[11] Pearson, K., *Tables of the Incomplete Beta-Function* (Cambridge University Press, Cambridge Massachusetts, 1934).

[12] Shephard, R. W., Lecture Notes on Inventory Theory, University of California, Berkeley (1966).

[13] Ravindran, A., "An Inventory Model with Contagious Demand Distribution," *ORC Report 69–19*, Operations Research Center, University of California, Berkeley (1969). (Unpublished Ph.D. dissertation.)

[14] Taylor, C. J., "The Application of Negative Binomial Distribution to Stock Control Problems," Operational Research Quarterly **12**, 81 (1961).

[15] Williamson, E. and M. H. Bretherton, *Tables of Negative Binomial Probability Distribution* (John Wiley and Sons, Inc., New York, 1963).

☆U.S. GOVERNMENT PRINTING OFFICE:1972-515-457/4

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

MARCH 1972
VOL. 19, NO. 1
NAVSO P-1278

CONTENTS

ARTICLES | Page

OFFICE OF NAVAL RESEARCH